JURNAL TIMES
TECHNOLOGY INFORMATICS & COMPUTER SYSTEM

Volume XI No 1, Juni 2022
*p*ISSN : 2337 – 3601
*e*ISSN : 2549 – 015X

Tersedia *online* di http://ejournal.stmik-time.ac.id

MODEL KLASIFIKASI GENETIC-XGBOOST DENGAN T-DISTRIBUTED STOCHASTIC NEIGHBOR EMBEDDING PADA PERAMALAN PASAR

Rimbun Siringoringo[1], Resianta Perangin Angin[2], Benget Rumahorbo[3]
Program Studi Teknik Informatika
Universitas Methodist Indonesia Medan
Jl. Hang Tuah No.8 Medan
email : rimbun.ringo@gmail.com [1], resianta88@gmail.com [2], benget888@gmail.com [3]

Abstrak

Extreme Gradient Boosting atau XGBoost merupakan metode *ensemble boosting* yang sangat populer dan berkinerja baik. Disisi lain XGBoost menerapkan sangat banyak parameter atau *hyper parameter*. Penentuan nilai secara manual tentu saja sangat sulit dan lama. Pada penelitian ini, Genetic Algoritm (GA) diterapkan untuk penelusuran nilai parameter XGBoost. Model XGBoost dievalusi dengan membandingkan ROC dengan beberapa model berbasis *tree*. Hasil pengujian ROC Genetic-XGBoost, Gradient Boost, dan Random Forest masing-masing sebesar 0, 987, 0,99, dan 0,957. Hasil ROC ke tiga model menunjukkan bahwa model Genetic-XGBoost memiliki performa yang lebih baik dari model-model lain.

Kata Kunci: XGBoost, Data mining, Genetic Algoritm, Genetic-XGBoost

1. Pendahuluan

Kondisi *outlier* merupakan munculnya nilai-nilai parameter yang ekstrim pada sebuah data. *Outlier* dapat berdampak pada kesalahan klasifikasi model, bias dalam memperkirakan parameter, hasil yang salah, dan prakiraan yang buruk [1]. Pada umumnya metode pengklasifikasi terbimbing tidak dilengkapi dengan kemampuan bawaan untuk menangani outlier, sehingga jika bekerja pada data outlier akan mengurangi kemampuan generalisasi ().

Teknik-teknik *ensemble* merupakan pendekatan baru dalam meningkatkan performa klasifikasi. Teknik ini dibangun dengan mengkombinasikan banyak basis pengklasifikasi. Teknik *ensemble* terbukti lebih baik jika dibandingkan dengan pengklasifikasi tunggal. Sejauh ini ada tiga pendekatan pada teknik ensemble, yaitu *bagging*, *boosting*, dan *stacking*.

Extreme Gradien Boosting atau XGBoost merupakan pengklasifikasi berbasis boosting. Dibandingkan dengan pendahulunya seperti *gradient boosting*, XGBoost memiliki kemampuan konvergensi serta generalisasi yang sangat baik [2], [3]. Hal tersebut terbukti dari performa metrik akurasi yang tinggi [4]. Selain hal tersebut, XGBoost memiliki kemampuan yang baik dalam mengolah data tidak seimbang. XGBoost memiliki kemamuan dalam membedakan fitur terpenting pada suatu data.

Selain kelebihan-kelebihan di atas, XGBoost adalah metode pengklasifikasi dengan banyak parameter atau *hyper parameter*. Penentuan nilai setiap parameter tergolong sulit mengakibatkan hasil yang diperoleh terjebak pada situasi *local optimum* [5]. Penentuan nilai setiap parameter secara manual tentu saja menghabiskan waktu yang tidak sedikit.[6]

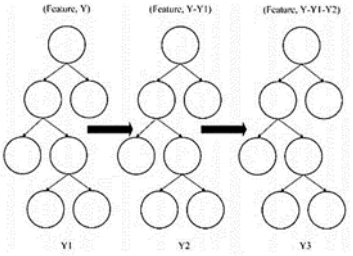
Pendekatan populer yang diterapkan pada penentua nilai parameter XGBoost adalah *grid search*, *random search*, dan *bayesian optimization* (Janizadeh et al., 2021; Y. Xia et al., 2017; Putatunda & Rama, 2019). Meskipun mudah untuk di implementasikan, namun metode-metode esebut bersifat *comptational expensive* [9].

Genetic Algorithm (GA) merupakan solusi penyelesaian masalah optimasi *hyper parameter* yang sangat populer. Keberhasilan GA pada tuning parameter telah teruji melalui berbagai penelitian. Beberapa diantaranya adalah pada penyelesaian *hyper parameter* apada aplikasi polusi air [10], dan penyelesaian *feedforward neural networks* [11]. Pada penelitian ini, akan dibangun kerangka klasifikasi dengan dua pendekatan yaitu penanganan *outlier* dan tuning *hyperparameter* XGBoost melalui GA.

2. Landasan Teori

XGBoost

XGBoost adalah teknik ensembel boosting berbasis pohon keputusan atau pohon regresi (gambar 1) [2]. Dengan menggabungkan beberapa pohon keputusan, nilai galat atau error diminimalisasi dari pohon pertama (feature, Y) ke pohon selanjutnya (feature, |Y-Y1|).



Gambar 1. Pohon regresi XGBoost

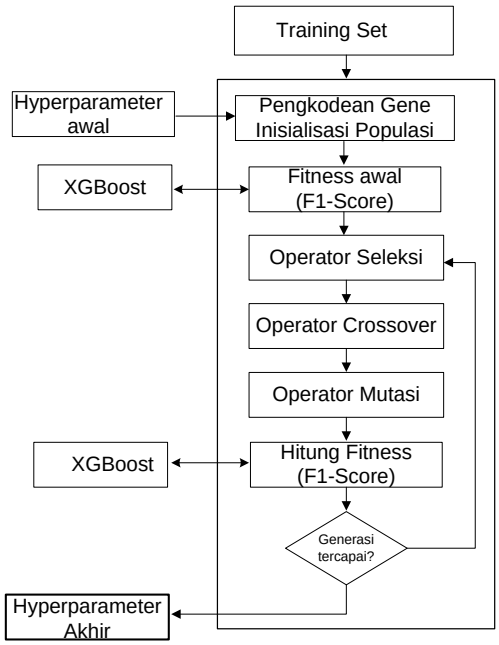
Interquartile Range

Interquartile Range (IQR) merupakan metode deteksi *outlier* berdasarkan nilai *quartile* (Q). Terdapat tiga jenis *quartile* yaitu Q_1, Q_2, dan Q_3. Q_1 adalah nilai yang merepresentasikan data diantara nilai terkecil dengan median Q_2 adalah nilai tengan data atau median, Q_3 adalah nilai yang merepresentasikan median dengan data terbesar. IQR dapat dihitung dengan persamaan (1)

$$IQR = Q_3 - Q_1 \quad (1)$$

3. Metode Penelitian

Gambaran umum prosedur dan tahapan penyelesaian masalah segmentasi pasar dengan model Genetic-XGBoost digambarkan pada gambar 2



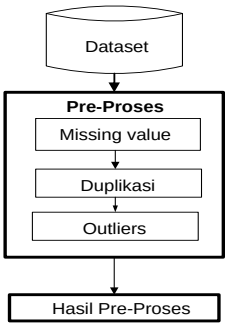
Gambar 2. Model Genetic-XGBoost

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

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.



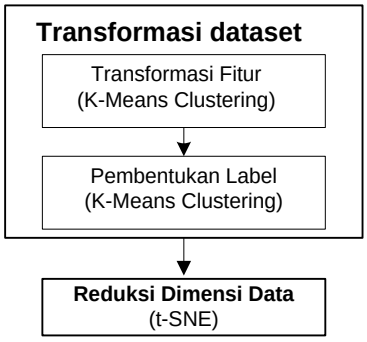
Gambar 3. Tahapan pre-proses

Model Genetic-XGBoost merupakan tenik pembelajaran tak terbimbing yang mengolah data berlabel atau memiliki kelas. Oleh karena itu perlu dilakukan transformasi data awal menjadi data yang diberi label atau klaster (gambar 4). Fitur data dengan dimensi *mn* dengan *m* adalah label atau klaster data, dan *n* merupakan banyak fitur data.

	fitur-1	fitur-2	…	fitur-n
label-1			…	
label-2			…	
label-3			…	
….	…	…	…	…
label-m			…	

Gambar 4. Model data

Diagram alur transformasi data data didasarkan pada gambar 5. Transformasi fitur bertujuan utuk membentuk fitur bari dataset yaitu dengan klaster fitur produk (*description*) menjadi beberapa kategori menggunakan *k-means clustering* (KMC). Selanjutnya adalah pembentukan label target dataset ke dalam beberapa kalster.

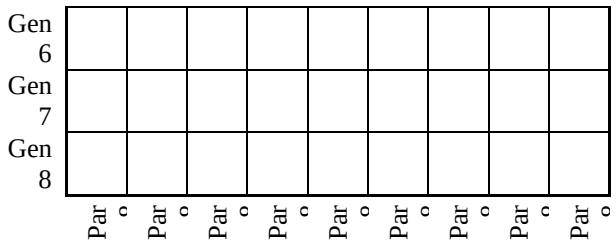


Gambar 5. Tahapan transformasi dataset

Skema pengkodean populasi

Pengkodean parameter XGBoost pada GA menerapkan matrik bilangan riil secara acak *uniform* dan dapat digambarkan sebagai berikut (gambar 5). Skema pengkodean GA tersebut membetuk matriks numerik berdimensi 8x8.

Gen 1									
Gen 2									
Gen 3									
Gen 4									
Gen 5									



Gambar 6. Skema pengkodean GA

Inisialisasi populasi

Pada tahap inisialisasi, terdapat tujuh parameter XGBoost yang dioptimasi. Nilai minimum dan maksimum yang diperbolehkan ditampilkan pada tabel 1.

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

Seleksi, Crossover, Mutasi

Operator seleksi menerapkan metode probabilitas acak dengan syarat (P_c>0.5) Metode *crossover* yang diterapkan adalah *uniform crossover*, mutasi dilakukan secara acak dengan rentang nilai yang diizinkan sesai dengan tabel 1.

4. Hasil Penelitian

Pra-Proses Data

Dataset mengandung duplikasi sebesar 445 data serta null value di atas 25%. Keberadaan duplikasi data dan null value dihapus dari dataset. Selanjutnya adalah tranformasi dataset awal (tanpa label) menjadi dataset baru (ber label). Transformasi dataset dilakukan melalui proses kategorissi produk dengan metode k-means clustering. Penentuan julah klaster ditentukan berdasarkan nilai silhouette score. Jumlah klaster terbaik adalah lima kategori yang diberi label {k_0, k_1, k_2, k_3, k_4}.

Hasil visualisasi klaster menggunakan t-Sne

Penerapan t-SNE 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

Setelah proses kategorisasi produk, dilakukan proses klasterisasi untuk menentukan label dataset. Penentuan label dilakukan dengan metode *k-means clustering* dengan jumlah klaster sebanyak 11. Pada tabel 3 ditampilkan sepuluh baris pertama hasil transformasi dataset yang terdiri dari lima fitur (k_0, k_1, k_2, k_3, k_4) dan sebelas kelas (0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10)

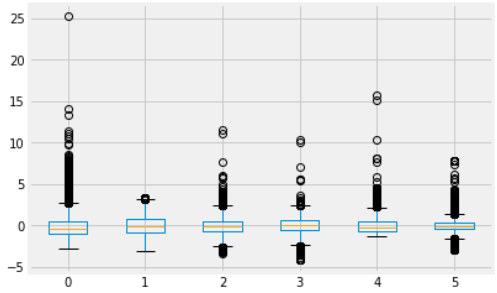
Tabel 3. Data hasil klaster pelanggan

No	k_0	k_1	k_2	k_3	k_4	cluster

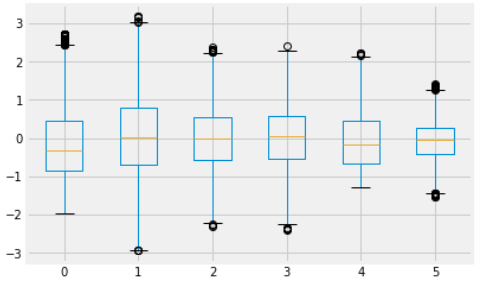
1	-1.782	-0.076	0.058	-0.226	-0.243	3
2	0.652	0.988	-2.501	-0.416	-0.605	0
3	-0.550	0.527	0.902	-0.734	0.708	3
4	-1.445	1.457	-2.071	1.138	-1.771	2
5	3.706	-0.280	0.165	-0.936	-0.376	1
6	-0.187	1.208	2.139	-0.592	-1.132	6
7	-0.008	-0.073	-0.645	0.149	-0.362	3
8	1.404	-0.213	-0.256	-0.513	-0.340	1
9	2.701	0.630	0.762	0.197	0.020	1
10	2.309	0.470	-0.374	0.097	-0.619	1

Pemrosesan outlier

Walaupun klaster data dapat terbentuk dengan baik, hasil klaster masih mengandung data outlier. Pemrosesan lanjutan dilakukan untuk mengetahui sebaran data outlier pada dataset. Pada gambar 8 ditampilkan visualisasi oulier dataset Pemrosesan awal diperlukan untuk meminimalisasi outlier. Penanganan outlier dilakukan dengan metode Inter Quartile Rank (IQR). Hasil penanganan outlier (gambar 9) menunjukkan bahwa outlier data dapat diminimalisasi.



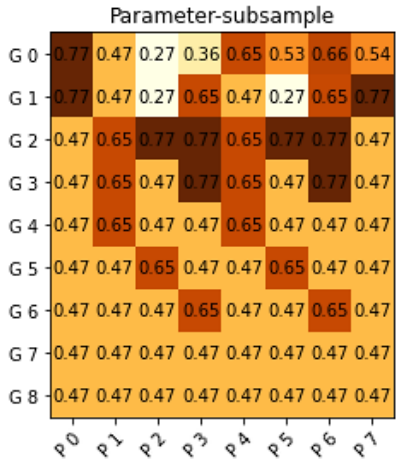
Gambar 8. Gambaran *outlier* dataset



Gambar 9. Gambaran data dengan *outlier* kecil

Tuning *hyperparameter* dengn Genetic-Xgboost

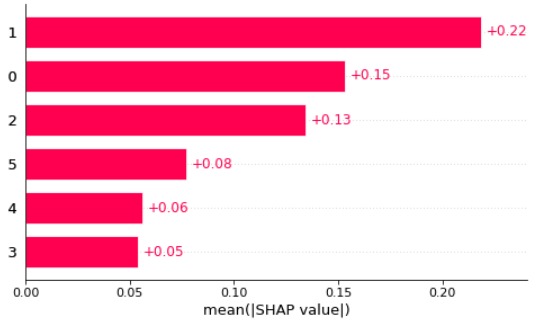
Dengan 500 iterasi, pada iterasi 100 diperoleh nilai parameter *learning_rate* 0.47, *n_estimators* 79.0, *max_depth* 3, *min_child_weight* 4.27, *gamma* 2.56 *sub-sample* 0.47, *colsample_bytree* 0.21. Selanjutnya pada gambar 10 ditampilkan grafik *heatmap* penelusuran solusi untuk parameter subsample. Melalui gambar tersebut diperoleh hasil pencarian solusi dengan nilai 0.47.



Gambar 10. *Heatmap* pencarian solusi

Fitur relevan

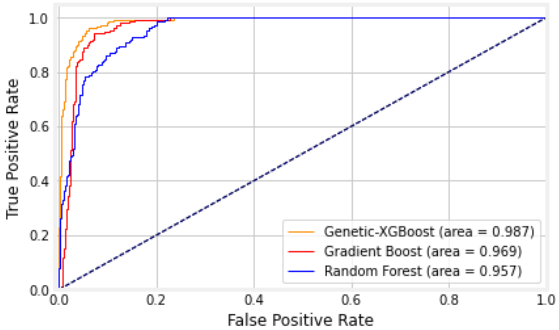
Pada gambar berikut ini ditampilkan grafik peringat fitur terpenting. Pada gambar tersebut, fitur 1 (k_0) menempati fitur dengan skor tertinggi yaitu +0.22



Gambar 11. Fitur penting dataset

Evaluasi model dengan Grafik ROC

Evalusi model didasarkan ada grafik ROC dengan membandingkan hasil ROC beberapa model berbasis *tree*.Hasil perbandingan grafik ROC ditampilkan pada gambar 9 berikut. ROC model Genetic-XGBoost, Gradient Boost, dan Random Forest masing-masing sebesar 0, 987, 0,99, dan 0,957. Hasil ROC ke tiga model menunjukkan bahwa model Genetic-XGBoost memiliki performa yang lebih baik dari model-model lain.



Gambar 12. Grafik ROC beberapa model classifier

5. 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

6. Daftar Pustaka

[1] L. Sunitha, D. M. BalRaju, and J. S. Kiran, “Detection and Analysis of Outliers and Applying Data Mining Methods on Weather Data of Bhanur Village Detection and Analysis of Outliers and Applying Data Mining Methods on Weather Data of Bhanur Village Abstract :,” no. January, 2021.

[2] Y. Jiang, G. Tong, H. Yin, and N. Xiong, "A Pedestrian Detection Method Based on Genetic Algorithm for Optimize XGBoost Training Parameters," *IEEE Access*, vol. 7, pp. 118310–118321, 2019, doi: 10.1109/access.2019.2936454.

[3] Y. Wang and Y. Guo, "Forecasting method of stock market volatility in time series data based on mixed model of ARIMA and XGBoost," *China Commun.*, vol. 17, no. 3, pp. 205–221, 2020, doi: 10.23919/JCC.2020.03.017.

[4] D. Wu, P. Guo, and P. Wang, "Malware Detection based on Cascading XGBoost and Cost Sensitive," in *2020 International Conference on Computer Communication and Network Security (CCNS)*, 2020, pp. 201–205. doi: 10.1109/CCNS50731.2020.00051.

[5] Y. Mai, Z. Sheng, H. Shi, and Q. Liao, "Using Improved XGBoost Algorithm to Obtain Modified Atmospheric Refractive Index," *Int. J. Antennas Propag.*, vol. 2021, p. 5506599, 2021, doi: 10.1155/2021/5506599.

[6] P. Wulandari and R. Gultom, "Pengaruh Likuiditas, Aktivitas, dan Pertumbuhan Penjualan Terhadap Profitabilitas Pada Perusahaan Industri Makanan dan Minuman Yang Terdaftar Di Bursa Efek Indonesia Tahun 2014-2017," *J. Ilm. Methonomi*, vol. 4, no. 2, pp. 101–110, 2018.

[7] S. Janizadeh, M. Vafakhah, Z. Kapelan, and N. Mobarghaee Dinan, "Hybrid XGboost model with various Bayesian hyperparameter optimization algorithms for flood hazard susceptibility modeling," *Geocarto Int.*, pp. 1–20, Oct. 2021, doi: 10.1080/10106049.2021.1996641.

[8] Y. Xia, C. Liu, Y. Li, and N. Liu, "A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring," *Expert Syst. Appl.*, vol. 78, pp. 225–241, 2017, doi: https://doi.org/10.1016/j.eswa.2017.02.017.

[9] S. Putatunda and K. Rama, "A Modified Bayesian Optimization based Hyper-Parameter Tuning Approach for Extreme Gradient Boosting," in *2019 Fifteenth International Conference on Information Processing (ICINPRO)*, 2019, pp. 1–6. doi: 10.1109/ICInPro47689.2019.9092025.

[10] X. Xia, S. Jiang, N. Zhou, X. Li, and L. Wang, "Genetic algorithm hyper-parameter optimization using Taguchi design for groundwater pollution source identification," *Water Supply*, vol. 19, no. 1, pp. 137–146, Mar. 2018, doi: 10.2166/ws.2018.059.

[11] C. Boonthanawat and C. Boonyasiriwat, "Finding optimal hyperparameters of feedforward neural networks for solving differential equations using a genetic algorithm," *J. Phys. Conf. Ser.*, vol. 1719, no. 1, p. 12033, 2021, doi: 10.1088/1742-6596/1719/1/012033.