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

SEGMENTASI DAN PERAMALAN PASAR RETAIL MENGGUNAKAN XGBOOST DAN PRINCIPAL COMPONENT ANALYSIS

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

DOI: https://doi.org/10.46880/jmika.Vol5No1.pp42-47

ABSTRACT

The growth of the online retail market in Indonesia is an excellent business opportunity. It is predicted that this growth will continue to move upward due to the increasing internet penetration. With greater exposure to brands, products and offerings, consumers become smarter and wiser in their purchasing decisions. Offering goods and services that match the tastes and behavior of consumers is very important to maintain business continuity. So far, the models developed are divided into two major parts, namely the time series approach and machine learning. In this study, segmentation and forecasting of online retail sector sales were carried out using extreme gradient boosting (XGBoost). The data used in this study is an online retail dataset obtained from the UCI repository. The k-means clustering (KMC) method is applied to determine the target or data class. Principal component analysis (PCA) is applied to reduce data dimensions by eliminating irrelevant features. Model evaluation is based on confusion matrix and macro average ROC curve. Based on the research results, XGBoost can perform retail data classification well, this can be seen through confusion matrix metrics and ROC curves.

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

ABSTRAK

Pertumbuhan pasar ritel *online* di Indonesia merupakan peluang bisnis yang sangat baik. Pertumbuhan ini diprediksi akan terus bergerak naik sehubungan dengan meningkatnya penetrasi internet. Dengan lebih masifnya pemaparan terhadap berbagai merek, produk, dan penawaran, konsumen menjadi lebih cerdas dan bijaksana dalam keputusan pembelian mereka. Penawaran barang dan jasa yang cocok dengan selera dan perilaku konsumen sangat penting untuk menjaga kelangsungan bisnis. Sejauh ini model model yang dikembangkan tebagi atas dua bagian besar yaitu pendekatan time series dan *machine learning*. Pada penelitian ini, dilakukan segmentasi dan peramalan terhadap penjualan sektor ritel *online* menggunakan *extreme gradient boosting* (XGBoost). Data yang digunakan pada penelitian ini adalah *online* retail dataset yang diperoleh dari repositori UCI. Metode *k-means clustering* (KMC) diterapkan untuk menentukan target atau kelas data. Principal componen analysis (PCA) diterapkan untuk mereduksi dimensi data dengan menghilangkan fitur yang tidak relevan. Evaluasi model didasarkan pada confusion matrix dan *macro average* ROC *curve*. Berdasarkan hasil penelitian, XGBoost dapat melakukan klasifikasi data retail dengan baik, hal tersebut terlihat melalui c*onfusion matrix* dan kurva ROC.

Kata Kunci: Extreme Gradient Boosting*, Peramalan Pasar, Segmentasi Konsumen, *Principal Component Analysis.

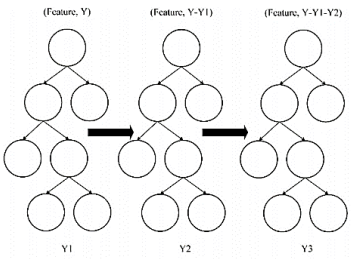
PENDAHULUAN

Indonesia merupakan negara dengan jumlah *mini-market* terbesar di kawasan Asia Tenggara. Pertumbuhan ritel di Indonesia meningkat sebesar 130 % pada rentang tahun 2010 sampai 2015 (Mulya, Si, Hermawan, & Evienia, 2019) dengan 43.826 toko pada pertengahan tahun 2019. Banyak faktor pendorong pertumbuhan tersebut, diantaranya adalah pertumbuhan usia muda kelas menengah (Aribawa, 2016) dan kehidupan yang semakin sibuk, terutama di kota-kota besar, lokasi yang mudah dijangkau, dan kenyamanan berbelanja menjadi sangat penting bagi konsumen. meningkatnya penetrasi internet dan lebih banyaknya pemaparan terhadap berbagai merek, produk, dan penawaran, konsumen di seluruh wilayah ini menjadi lebih cerdas dan bijaksana dalam keputusan pembelian mereka. Penawaran yang disesuaikan dengan selera lokal dan perilaku pembelian sangat penting untuk menjaga kelangsungan bisnis. Pada area ini, sistem-sistem cerdas dan *data science* dapat membantu penyelesaian masalah dari aspek segmentasi pasar dan konsumen.

TINJAUAN PUSTAKA

XGBoost

Metode *extreme gradient boosting* atau XGBoost merupakan sebuah algoritma boosting berbasis pohon keputusan atau pohon regresi. Gambaran umum algoritma *boosting* berbasis pohon regresi ditampilkan pada gambar 1 (Jiang, Tong, Yin, & Xiong, 2019). Proses pembelajaran pohon pertama dari data latih (*feature*, Y) memperoleh hasil estimasi pertama (Y1). Pohon ke dua melakukan proses pembelajaran dari data latih (*feature*, |Y-Y1|), dimana nilai |Y-Y1| merupakan selisih antara label nyata dengan label prediksi pada tahap sebelumnya. Pohon ketiga melakukan proses pembelajaran dari data (*feature*, |Y-Y1-Y2|) dan menghasilkan estimasi Y3. Dengan cara tersebut, nilai *error* dapat direduksi dengan efektif.



Gambar 1. Pohon Regresi XGBoost

Principal Component Analysis

Metode *principal component analysis* (PCA) adalah metode reduksi dimensi yang sering digunakan untuk mengurangi dimensi data yang besar. PCA mampu mengubah sekumpulan besar variabel menjadi kelompok data yang lebih kecil yang masih berisi sebagian besar informasi dalam kumpulan besar. Sebuah matrik *X* berukuran *n* atribut × *m* data, dapat di dekomposisi ke dalam *m* buah vektor. *X* dapat diekspresikan pada persamaan (1) (Zhang, Zhang, & Wu, 2020). Dimana $ti \in R^n$ merupakan nilai vektor

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

Selanjutnya, nilai vektor dari *X* disebut sebagai principal component dari matrik *X*. Jika beberapa faktor minor diabaikan dan hanya sebanyak *a* elemen *principal* yang diterapkan, makan matrik *X* dapat diekspresikan pada persamaan (2).

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

Penelitian terkait

Terdapat banyak penelitian di bidang peramalan dengan pendekatan yang sangat beragam. Sejauh ini model model yang dikembangkan di fokuskan pada dua aspek yaitu pendekatan *time series* (Anggraeni, Andri, Sumaryanto, & Mahananto, 2017), dan *machine learning* (Chatzis, Siakoulis, Petropoulos, Stavroulakis, & Vlachogiannakis, 2018). Pendekatan *time series* dilakukan dengan menganalisis pola perubahan karakteristik data terhadap waktu. Pendekatan *machine learning* merupaka bagian dari kecerdasan buatan dengan belajar dari data.

XGBoost merupakan metode *machine learning* yang sangat populer saat ini. Algoritma XGBoost merupakan salah satu teknik yang populer pada kelompok *boosting* karena memiliki konvergensi yang sangat baik (Jiang et al., 2019). Terdapat banyak penelitian yang berkaitan dengan Pramalan dan segmentasi pasar menggunakan XGBoost. Diantaranya (Wang & Guo, 2020) melakukan peramalan terhadap stok beras. Model berbasis XGBoost tersebut berhasil memprediksi pembukaan harga awal dengan sangat baik. XGBoost dapat diterapkan pada prediksi loyalitas pelanggan atau *churn* (Yayun, 2018) dengan akurasi yang tinggi. Pada bidang peramalan penjualan (Ji, Wang, Zhao, & Guo, 2019) menerapkan XGBoost pada platform *e-commerce.* XGBoost dapan memberikan hasil yang lebih baik dibanding metode-metode yang lain. XGBoost pada peramalan harga saham [12] menghasilkan tingkat generalisasi yang baik sehingga mampu memprediksi harga saham pembukaan dengan tepat. Penerapan XGBoost pada deteksi malware (Wu, Guo, & Wang, 2020) menghasilkan tingkat akurasi yang tinggi.

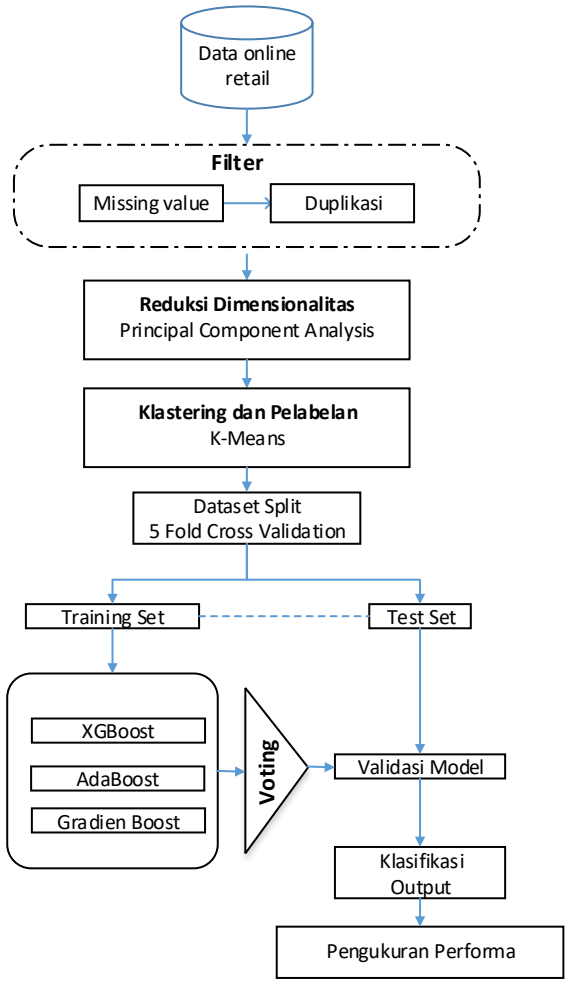
Ketika bekerja pada data berdimensi tinggi, banyak metode *machine learning* tidak menghasilkan performa yang baik. Hal tersebut ditunjukkan melalui beberapa penelitian terkait diantaranya adalah (Salim & Mitton, 2020) menerapkan *Machine Learning Based Data Reduction Algorithm* (MLDR) pada data sensor untuk pemantauan pertanian dimana proses reduksi data meningkatkan akurasi *machine learning.* Penerapan reduksi data sensor (Radhika & Rangarajan, 2019) berhasil meningkatkan kinerja *machine learning* dan efisiensi sensor.

Pada penelitian ini, dilakukan segmentasi dan peramalan terhadap penjualan sektor *ritel online* menggunakan *extreme gradient boost* (XGBoost). Untuk menujang kinerja XGBoost, metode *principal component analysis* (PCA) diterapkan guna mereduski dimensionalitas data ritel.

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

METODOLOGI

Kerangka kerja metode yang diusulkan pada penelitian ini dapat digambarkan pada gambar 2.



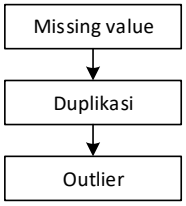
Gambar 2. Model yang diusulkan

Data awal yang digunakan pada penelitian ini adalah diperoleh dari repositori UCI yaitu *online retail dataset* (Chen, Sain, & Guo, 2012) dengan 525.461 item data dan 8 atribut.

Tabel 1. Kategori fitur

Atribut	**Unit**	**deskripsi**
InvoiceNo	Nominal	Nomor unik setiap transaksi
StockCode	Nominal	Kode produk
Description	Nominal	Nama produk
Quantity	Numeric	Banyak produk per transaksi
InvoiceDate	Numeric	Tanggal transaksi
UnitPrice	Numeric	Harga produk
CustomerID	Nominal	Nomor unik setiap pelanggan
Country	Nominal	Negara pelanggan

Pemrosesan awal bertujuan untuk mengatasi *missing value*, *outlier* dan duplikasi di dalam dataset. Atribut CustomerID memiliki *missing value* atau *null value* sebesar 20%. Nilai ini termasuk nilai yang besar dan oleh karenanya perlu di hapus dari dataset. Alur pra-proses data ditunjukkan pada gambar 3.



Gambar 3. Alur pra-proses data

Dataset *online retail* merupakan data dengan dimensionalitas yang tinggi. Terdapat ratusan ribu *record* data yang tersebar pada 8 atribut. Data dengan dimensionalitas yang tinggi menghasilkan sebaran data yang sangat padat sehingga model yang dihasilkan akan sangat sulit menghasikan akurasi yang baik. Algoritma *principal componen analysis* (PCA) diterapkan untuk mereduksi dimensi data dengan menghilangkan fitur yang tidak relevan.

Transformasi data bertujuan untuk menyediakan data berdimensi *mxn*, dimana *m* adalah target label, dan *n* adalah fitur data. Fitur data (fitur 1, fitur 2, fitur 3… fitur *n*) ditentukan melalui proses klaster produk, dan target label data (target 1, target 2, target 3…target *n*) ditentukan melalui klaster konsumen. Model tranformasi data ditampilkan pada gambar 4.

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

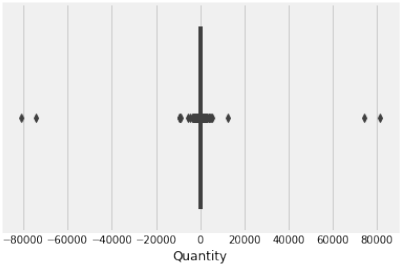
Gambar 4. Model dataframe berdimensi *m x n*

Metode *k-means clustering* (KMC) diaplikasikan untuk menentukan kelompok atau klaster data dimana jumlah klaster ditentukan berdasarkan nilai koefisien *Silhouette*. Sebelum proses klastering, dimensi dataset terlebih dahulu di reduksi menggunakan metode PCA. Klastering dilakukan terhadap produk maupun pelanggan. Pembagian sampel data training dan testing dilakukan secara acak dengan metode *5 fold cross validation*.

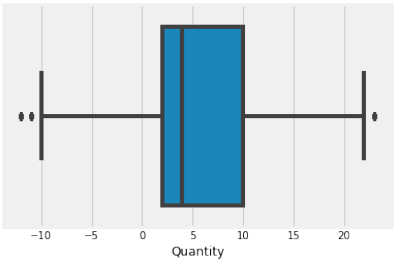
ISSN: 2598-8565 (media cetak)
ISSN: 2620-4339 (media online)

HASIL DAN PEMAHASAN

Atribut *UnitPrice* dan *Quantity* merupakan atribut yang mengandung *outlier* sebagaimana ditampilkan pada gambar 5. Dengan metode *Inter Quartile Range* (IQR), *outlier* data direduksi sehingga menyajikan data yang lebih baik sebagaimana ditampilkan pada gambar 6.

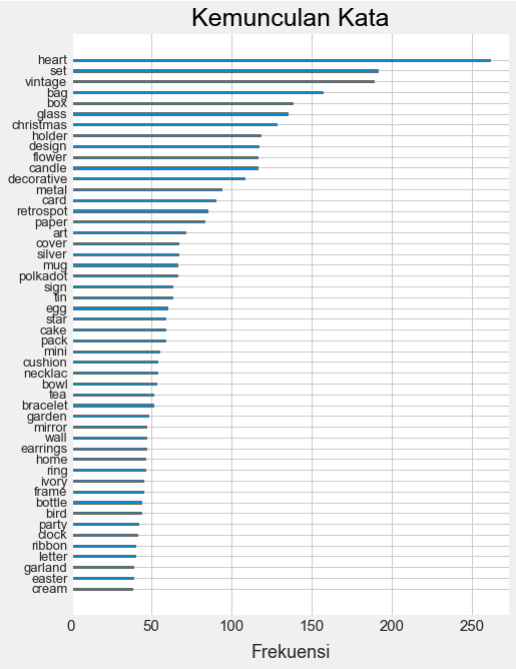


Gambar 5. *Boxplot outlier* dalam data



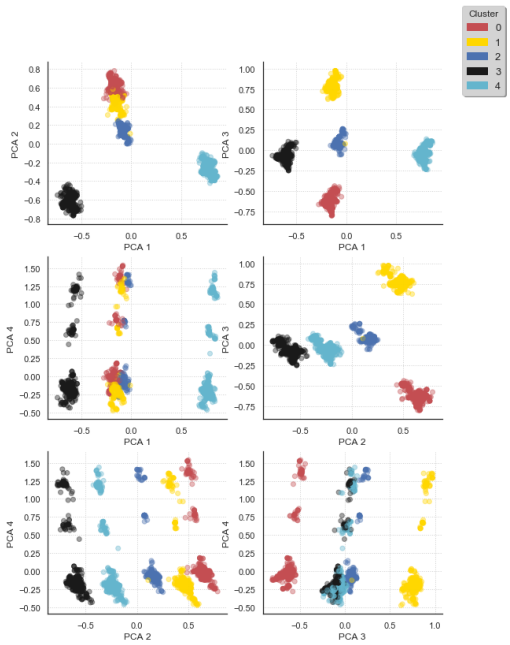
Gambar 6. *Boxplot* tanpa *outier*

Terdapat 1404 kata kunci atau *keyword* produk pada dataset. Frekuensi kemunculan setiap *keyword* ditampilkan pada gambar 7.



Gambar 7. Frekuensi kemunculan keyword

Untuk menyajikan hasil klaster yang elegan dimana setiap elemen benar benar terkelompokkan dengan baik, pada tahap selanjutnya diterapkan PCA dengan jumlah komponen 50 dan koefisien *alpha* 0.4. PCA dapat menyajikan hasil klaster yang lebih baik, hal tersebut terbukti melalui warna klaster yang terpisah satu sama lain seperti yang ditampilkan pada gambar 8.



Gambar 8. Visualisasi klaster

Penentuan fitur data ditentukan melalui proses klaster produk menggunakan KMC dan *hamming distance*. Penentuan jumlah klaster optimal ditentukan berdasarkan nilai koefisien *silhouette*. Koefisien *silhouette* untuk jumlah klaster n={3,4,5,6,7,8,9} adalah { 0.091, 0.118, 0.167, 0.137, 0.136, 0.147, 0.136}. Nilai maksimum diperoleh sebesar 0.167 sehingga jumlah klaster optimum = 5.

Pada tabel 2 berikut ditampilkan hasil klaster lima jenis produk. Klaster final didasarkan pada nilai *k* terbesar di antara *k_0* sampai *k_4*. Produk *white hanging* memiliki nilai *k* terbesar pada *k_0* sehingga produk tersebut berada pada klaster 0, demikian juga produk *white metal* memiliki *k* terbesar pada *k_1* sebesar 20.34, sehingga di kelompokkan pada klaster 1

Tabel 2. Klaster produk

Deskripsi	k_0	k_1	k_2	k_3	k_4	final
WHITE HANGING	**15.3**	0.00	0.0	0.0	0.0	1
WHITE METAL	0.0	**20.34**	0.0	0.0	0.0	2
CREAM CUPID	0.0	**22.00**	0.0	0.0	0.0	2

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

KNITTED UNION	0.0	**20.34**	0.0	0.0	0.0	2
RED WOOLLY	0.0	**20.34**	0.0	0.0	0.0	2

Label data atau *class* ditentukan melalui klaster data pelanggan menggunakan KMC. Dengan metode koefisien *silhouette* terbesar diperoleh pada jumlah klaster = 11. Terdapat 11 klaster pelanggan dan 5 klaster produk. Pada tabel 3 ditampilkan 10 sampel hasil klaster pelanggan. Berdasarkan tabel tersebut di peroleh gambaran bahwa produk dengan nomor urut 1 di segmentasikan kepada konsumen tipe klaster 3, demikian juga produk nomor urut 2 di segmentasikan kepada konsumen tipe klaster 0. hubungan antara klaster konsumen dengan klaster produk. Konsumen klaster 1 merupakan klaster dengan jumlah konsumen (jlh_kons) terbanyak yaitu 1020 konsumen. Konsumen klaster 9 merupakan klaster dengan total pembelian (total) terbesar yaitu 3313,012. *Data frame* yang diterapkan pada klasifikasi adalah X={k_0, k_1, k_2, k_3, k_4} dan target label data adalah Y={klaster} Dari sisi jumlah transaksi per *user* (jlh_trans), klaster konsumen 9 merupakan klaster konsumen dengan jumlah transaksi per *user* terbesar yaitu 86 transaksi

Tabel 3. Klaster Pelanggan

No	k_0	k_1	k_2	k_3	k_4	cluster
1	-1.781	-0.075	0.057	-0.225	-0.242	3
2	0.651	0.987	-2.500	-0.415	-0.604	0
3	-0.549	0.526	0.901	-0.733	0.707	3
4	-1.444	1.457	-2.070	1.137	-1.770	2
5	3.705	-0.279	0.164	-0.935	-0.375	1
6	-0.186	1.207	2.138	-0.591	-1.131	6
7	-0.007	-0.072	-0.644	0.148	-0.361	3
8	1.403	-0.212	-0.255	-0.512	-0.339	1
9	2.700	0.629	0.761	0.196	0.019	1
10	2.308	0.469	-0.373	0.096	-0.618	1

Pada tabel 4 berikut dijabarkan parameter algortima XGBoost.

Tabel 4. *Hyperparameter* XGBoost

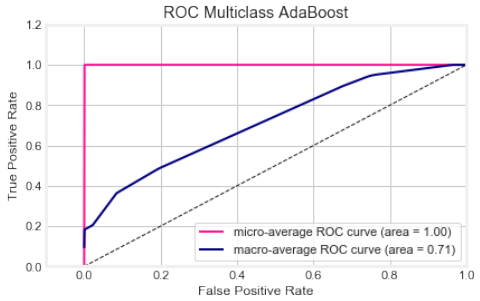
Hyperparameter	Nilai
max_depth	6
min_child_weight	1
base_score	0,5
gamma	0
reg_alpha	0
learning_rate	0,30

Pada tahap *training*, data hasil klaster pada pembahasan sebelumnya akan diterapkan pada proses klasifikasi. Split data set diatur dengan perbandingan (20 : 80) % serta menerapkan *5-fold cross validation.* Evaluasi model didasarkan pada kriteria-kriteria yang sudah populer yakni *confusion matrix, accuracy, precision, recall, f1* dan kurva ROC.

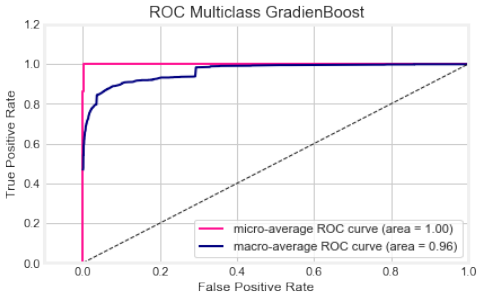
Tabel 5. Evaluasi Klasifikasi Training

	Accuracy	Precision	Recall	F1
Training	0.87	0.78	0.77	0.77
Testing	0.89	0.79	0.80	0.80
Rata-rata	0.88	0.78	0.78	0.78

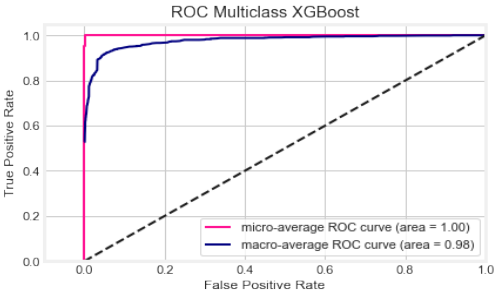
Kurva *receiver operating characteristics* (ROC) menggambarkan keakuratan model klasifikasi. Pada gambar 9 berikut disajikan grafik ROC untuk model-model klasifikasi. Kriteria ROC *multiclass* ditentukan melalui rata-rata makro atau *macro average* ROC *curve*. Pada gambar 9 berikut ditampilkan hasil perbandingan kurva ROC untuk beberapa model *boosting* untuk XGBoost, Gradien Boost, dan AdaBoost. Masing-masing nilai ROC adalah sebesar 0,98; 0,96; 0,71



(a). Kurva ROC AdaBoost



(b). Kurva ROC GradienBoost



(c). Kurva ROC XGBoost

Gambar 9. Kurva ROC beberapa metode *boosting*

Dari ke tiga grafik ROC di atas, XGBoost memiliki macro-average yang lebih baik dari dua metode boosting lainnya.

KESIMPULAN

Pada penelitian ini, dilakukan segmentasi dan peramalan terhadap penjualan sektor ritel *online* menggunakan *extreme gradient boost* (XGBoost). Metode *principal component analysis* (PCA) diterapkan untuk mereduksi dimensi dataset. Menggunakan kriteria *silhoulette*, KMC dapat menentukan target data dengan pendekatan klaster yang lebih terpisah dengan baik. Berdasarkan hasil penelitian di atas diperoleh kesimpulan bahwa XGBoost dapat melakukan klasifikasi data retail dengan baik, hal tersebut terlihat melalui kriteria metrik dan grafik ROC.

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.

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

Chen, D., Sain, S. L., & Guo, K. (2012). Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. *Journal of Database Marketing & Customer Strategy Management*, *19*(3), 197–208. https://doi.org/10.1057/dbm.2012.17

Ji, S., Wang, X., Zhao, W., & Guo, D. (2019). An Application of a Three-Stage XGBoost-Based Model to Sales Forecasting of a Cross-Border E-Commerce Enterprise. *Mathematical Problems in Engineering*, *2019*, 8503252. https://doi.org/10.1155/2019/8503252

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

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

Radhika, S., & Rangarajan, P. (2019). On improving the lifespan of wireless sensor networks with fuzzy based clustering and machine learning based data reduction. *Applied Soft Computing*, *83*, 105610. https://doi.org/https://doi.org/10.1016/j.asoc.2019.105610

Salim, C., & Mitton, N. (2020). *Machine Learning Based Data Reduction in WSN for Smart Agriculture BT - Advanced Information Networking and Applications* (L. Barolli, F. Amato, F. Moscato, T. Enokido, & M. Takizawa, Eds.). Cham: Springer International Publishing.

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

Yayun, Z. (2018). Research on E-commerce Customer Churn Prediction Based on Improved Value Model and XG-Boost Algorithm. *Management Science and Engineering*, *12*(3), 51–56. https://doi.org/10.3968/10816

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