

## RESEARCH ARTICLE

# Implementasi Model XGBoost untuk Prediksi Jumlah Transaksi dan Total Pendapatan di Jaringan Restoran CV Balibul

Muhammad Garma Asyam Rianto, Anggunmeka Luhur Prasasti \* and Astri Novianty

Fakultas Teknik Elektro, Universitas Telkom, Bandung, 40257, Jawa Barat, Indonesia

\*Corresponding author: [anggunmeka@telkomuniversity.ac.id](mailto:anggunmeka@telkomuniversity.ac.id)

## Abstrak

Penelitian ini bertujuan untuk menerapkan model XGBoost dalam memprediksi jumlah transaksi dan total pendapatan di jaringan restoran CV Balibul. Model XGBoost menggunakan teknik *gradient tree boosting* untuk meningkatkan akurasi prediksi dari data transaksi harian yang diolah menggunakan *library Pandas*. Optimisasi parameter untuk model dilakukan dengan metode *Bayesian Optimization*, dan evaluasi model menggunakan metrik  $R^2$ , RMSE, MAPE, dan *Pattern Similarity*. Hasil penelitian menunjukkan bahwa model XGBoost dapat memprediksi jumlah transaksi dan total pendapatan dengan tingkat akurasi yang masuk akal, di mana *shift* 1 memiliki nilai *error* yang lebih kecil dibandingkan *shift* 2.

**Key words:** XGBoost, Prediksi Transaksi, Total Pendapatan, *Bayesian Optimization*, CV Balibul.

## Pendahuluan

Perkembangan teknologi pada zaman ini berkembang dengan pesat. Dalam zaman yang serba digital, usaha - usaha seperti restoran CV Balibul ingin melakukan digitalisasi terhadap usaha mereka. Salah satu bentuk digitalisasi ini adalah penerapan sistem prediksi terhadap jumlah transaksi dan total pendapatan. Penerapan sistem prediksi ini ditujukan untuk mengoptimalkan manajemen sumber daya dan perencanaan operasional. Dalam penelitian ini, sistem prediksi yang digunakan merupakan sistem prediksi dengan model XGBoost. XGBoost dipilih karena keunggulannya dalam menangani data yang kompleks dan mampu memberikan hasil prediksi yang lebih baik dibandingkan metode lain seperti *Random Forest*. Penggunaan metode *Bayesian Optimization* dapat mengatasi masalah terhadap penentuan nilai parameter yang dibutuhkan oleh model.

## Tinjauan Pustaka

### XGBoost untuk Model Prediksi

XGBoost merupakan sebuah model *Machine Learning* berbasis *tree*. Tidak seperti model *Random Forest* yang menggunakan konsep bagging *tree*, model XGBoost menggunakan konsep *gradient tree boosting*. Setiap literasi [Opt] training pada model ini, *tree* pada model ini akan di *boosting* berdasarkan dari nilai gradien yang diperoleh dari iterasi sebelumnya [1]. Dalam memperoleh nilai prediksi, model ini akan menjumlahkan nilai dari setiap *tree*-nya [2].

### Pandas Sebagai *Library* Pengolahan Data

Pandas merupakan sebuah *library Python* yang memiliki kemampuan untuk melakukan pengolahan data seperti melakukan penambahan

dan pengurangan atribut data dan agregasi data. Pandas dapat mengambil data dari CSV, TXT, Excel, maupun dari *database* seperti MySQL [3].

### *Bayesian Optimization* Sebagai Penentu Parameter Model

*Bayesian Optimization* merupakan sebuah metode untuk melakukan optimisasi parameter model *machine learning*. Metode ini memiliki dua konsep utama, yaitu *Gaussian Process* dan *Acquisition Function*. *Gaussian Process* berperan sebagai model pengganti (*surrogate model*) yang dapat memberikan distribusi prediktif (rata - rata dan variansi) terhadap titik pencarian. *Acquisition Function* berperan sebagai pemandu untuk menentukan titik pencarian selanjutnya untuk dievaluasi [4].

### $R^2$ , RMSE, MAPE, dan *Pattern Similarity* sebagai *Metric* Evaluasi Model

$R^2$  merupakan suatu *metric* yang dapat mengukur seberapa baik model memiliki kecocokan dengan data. Nilai  $R^2$  memiliki rentang dari negatif tak hingga sampai 1. Nilai terbaik dari  $R^2$  adalah 1, semakin mendekati 1 maka semakin baik model dapat mempresentasikan data. Dalam kata lain, nilai  $R^2$  memberikan gambaran sejauh mana model dapat menjelaskan variasi dalam data [5].

$$R^2 = 1 - \frac{\sum(y - \hat{y})^2}{\sum(y - \bar{y})^2} \quad (1)$$

RMSE (*Root Mean Squared Error*) merupakan suatu *metric* yang mengukur besar *error* yang dihasilkan antara hasil prediksi dari model

**Table 1.** Interpretasi Nilai MAPE

MAPE (%)	Interpretasi
< 10	Prediksi sangat akurat
10 – 20	Prediksi yang baik
20 – 50	Prediksi yang masuk akal
> 50	Prediksi tidak akurat

dengan data sebenarnya. Nilai RMSE memiliki rentang 0 sampai positif tak hingga. Semakin kecil nilai RMSE, maka hasil prediksi model mendekati data sebenarnya [6].

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

MAPE (*Mean Absolute Percentage Error*) menunjukkan seberapa baik model dapat melakukan prediksi jika dibandingkan dengan data sebenarnya. Nilai MAPE memiliki rentang 0 sampai positif tak hingga. Semakin kecil nilai MAPE, maka prediksi model semakin akurat [7].

$$MAPE = \frac{1}{m} \sum_{i=1}^m \left| \frac{Y_i - X_i}{Y_i} \right| \times 100\% \quad (3)$$

Nilai MAPE dapat diinterpretasikan pada tabel 1 berikut. *Pattern Similarity* menunjukkan tingkat kecocokan data yang dihasilkan oleh model dengan data sebenarnya. *Pattern Similarity* tidak bergantung kepada besar *error* yang diperoleh. Nilai *Pattern Similarity* diperoleh dengan basis dari nilai *Pearson Correlation Coefficient*.

$$pcc(X, Y) = \frac{\sum_{i=1}^m (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^m (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^m (Y_i - \bar{Y})^2}} \quad (4)$$

Nilai *Pearson Correlation Coefficient* memiliki rentang -1 sampai 1. Nilai -1 berarti data prediksi dari model memiliki pola yang berlawanan dengan data sebenarnya [8]. Untuk mendapatkan nilai *Pattern Similarity*, nilai *Pearson Correlation Coefficient* diubah kedalam bentuk persentase dengan anggapan bahwa nilai -1 berarti tidak ada kecocokan pola.

$$pattern\ similarity = \left( \frac{correlation + 1}{2} \right) \times 100 \quad (5)$$

## Metodologi Penelitian

Bagian ini akan menjelaskan metode - metode yang digunakan untuk penelitian ini.

### Pengolahan Dataset

Dataset yang digunakan untuk pembuatan model ini berasal dari Qasir. Data yang berasal dari Qasir merupakan data item per transaksi, dalam arti lain data ini merupakan data yang berisi produk yang dibeli dari restoran. Gambar 1 Data Transaksi pada Halaman Qasir Dataset yang berasal dari Qasir dapat diunduh dengan format Excel. Data Excel ini terdiri dari dua puluh dua atribut yang dapat dilihat pada tabel 2. Dari atribut tersebut, tidak semua atribut digunakan untuk pembuatan model XGBoost. Data yang dibutuhkan merupakan data transaksi per hari. Dalam satu hari ini, terdapat dua *shift* kerja, *shift* pertama memiliki rentang dari jam delapan pagi sampai sebelum jam dua siang, *shift* kedua memiliki rentang dari jam dua siang sampai sebelum jam sepuluh malam. Pengolahan data item menjadi data transaksi per hari dapat dilakukan dengan menggunakan metode agregasi. Agregasi

**Gambar 1.** Data Transaksi pada Halaman Qasir.**Table 2.** Daftar Atribut pada Data Qasir

No.	Atribut
1.	No. Struk
2.	Tanggal
3.	Nama Outlet
4.	Nama Kasir
5.	Nama Pelanggan
6.	Produk
7.	Opsi Tambahan
8.	Jumlah Produk
9.	Jumlah Dibatalkan
10.	Harga Per Produk
11.	Subtotal
12.	Tipe Harga
13.	Diskon Produk
14.	Tipe Diskon Produk
15.	Disjon Transaksi
16.	Tipe Diskon Transaksi
17.	Redeem Poin
18.	Pajak
19.	Total
20.	Status
21.	Kode Pembayaran
22.	Metode Pembayaran
23.	No. Referensi

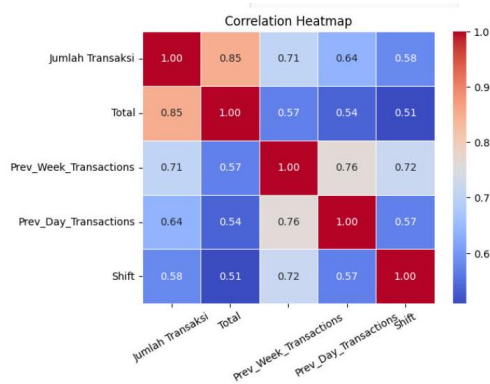
dilakukan berdasarkan atribut tanggal, no struk, dan *shift* dengan menjumlahkan atribut jumlah transaksi yang berasal dari no struk dan total.

### Penentuan Fitur Utama Prediksi

Data transaksi pada restoran ini memiliki pola data yang berbeda pada bulan yang sama. Oleh karena itu, fitur utama prediksi perlu ditentukan untuk menyesuaikan hasil prediksi yang dihasilkan oleh model. Fitur utama ini dapat diambil dari jumlah transaksi sebelumnya dari hari yang akan diprediksi. Data jumlah transaksi sebelumnya dapat diperoleh dari satu hari sebelumnya atau jumlah dari satu minggu sebelumnya. Fitur utama dapat ditentukan berdasarkan nilai korelasi antar atribut. Nilai korelasi dapat diperoleh dengan menggunakan metode Pearson dengan rentang nilai -1 sampai 1. Nilai Pearson -1 memiliki arti bahwa kedua atribut memiliki pola yang berlawanan [8].

	Tanggal	Shift	Jumlah Transaksi	Total
0	2020-09-18	1	7	962000
1	2020-09-18	2	34	5237000
2	2020-09-19	1	17	3162000
3	2020-09-19	2	21	3425000
4	2020-09-20	1	18	2345000
...	...	...	...	...
2730	2024-06-28	2	27	5851750
2731	2024-06-29	1	15	2686750
2732	2024-06-29	2	29	7346250
2733	2024-06-30	1	13	2452500
2734	2024-06-30	2	24	4242000

Gambar 2. Data Transaksi per Hari.



Gambar 3. Korelasi antar Atribut.

Table 3. Hasil R2 terhadap Jumlah Transaksi

Parameter	Batas Bawah	Batas Atas
<i>max_depth</i>	3	10
<i>learning_rate</i>	0,05	0,15
<i>n_estimators</i>	50	1000

Table 4. Nilai Parameter Terbaik

Parameter	Nilai	
	Shift 1	Shift 2
<i>max_depth</i>	3	6
<i>learning_rate</i>	0,1274	0,1297
<i>n_estimators</i>	564	812

$$pcc(X, Y) = \frac{\sum_{i=1}^m (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^m (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^m (Y_i - \bar{Y})^2}} \quad (6)$$

Atribut jumlah transaksi satu minggu sebelumnya memiliki nilai korelasi 3 yang lebih tinggi dari jumlah transaksi satu hari sebelumnya. Oleh karena itu, fitur ini yang dipilih menjadi fitur utama untuk prediksi.

### Model Machine Learning

XGBoost memiliki beberapa parameter yang dapat diatur untuk optimisasi model, diantaranya yaitu *n\_estimator*, *max\_depth*, dan *learning\_rate* ( $\eta$ ). Nilai untuk parameter parameter ini akan ditentukan

Table 5. Hasil R2 terhadap Jumlah Transaksi

Data	R <sup>2</sup> Jumlah Transaksi
Shift 1	0.2992
Shift 2	0.4211

Table 6. Hasil R2 terhadap Total Pendapatan

Data	R <sup>2</sup> Jumlah Transaksi
Shift 1	0.2507
Shift 2	0.5021

Table 7. Hasil RMSE Terhadap Jumlah Transaksi

Data	R <sup>2</sup> Jumlah Transaksi
Shift 1	4.07
Shift 2	6.79

Table 8. Hasil RMSE Terhadap Total Pendapatan

Data	R <sup>2</sup> Jumlah Transaksi
Shift 1	829392.17
Shift 2	1439253.98

Table 9. Hasil MAPE Terhadap Jumlah Transaksi

Data	MAPE Jumlah Transaksi
Shift 1	0.2344
Shift 2	0.2599

dengan menggunakan *Bayesian Optimization*. Iterasi dari pencarian nilai ini sebanyak dua puluh dengan lima iterasi awal sebelum pencarian. Batasan nilai parameter tersebut terletak pada tabel 3 ini.

Dari dua puluh lima iterasi total terhadap model, nilai parameter terbaik terletak pada iterasi terakhir untuk kedua model dengan perolehan nilai parameter yang berbeda antar model. Nilai parameter tersebut digunakan untuk fitting model sebenarnya terdapat pada tabel 4. Model XGBoost akan secara otomatis memilih iterasi *fitting* terbaik. Model *shift 1* memiliki iterasi terbaik pada iterasi ke-16, sedangkan model *shift 2* memiliki iterasi terbaik pada iterasi ke-52.

## Hasil dan Pembahasan

Model yang telah dibuat akan diuji dengan menggunakan empat *metric*, yaitu R<sup>2</sup>, RMSE, MAPE, dan *Pattern Similarity*. Data yang digunakan untuk pengujian merupakan data satu bulan setelah bulan terakhir data untuk *fitting* model secara terpisah berdasarkan *shift*.

### Pengujian dengan R<sup>2</sup>

R<sup>2</sup> menunjukkan tingkat kecocokan titik data antara prediksi dengan sebenarnya. Hasil pengujian R<sup>2</sup> terletak pada tabel 5 & 6 ini. Dari hasil R<sup>2</sup> yang diperoleh, model *shift 2* memiliki performa yang lebih baik dari *shift 1*.

### Pengujian dengan RMSE

RMSE menunjukkan seberapa besar *error* yang dihasilkan antara hasil prediksi dengan sebenarnya. Hasil pengujian RMSE terletak pada tabel 7 & 8 ini. Model *shift 1* dapat memprediksi nilai jumlah transaksi dan total pendapatan dengan *error* yang lebih kecil dari model *shift 2*.

### Pengujian dengan MAPE

MAPE menunjukkan apakah model dapat melakukan prediksi dengan akurat, masuk akal, ataupun buruk. Hasil pengujian MAPE terletak pada tabel 9 & 10 ini. Semua hasil MAPE pada tabel 9 & 10

**Table 10.** Hasil MAPE Terhadap Total Pendapatan

Data	MAPE	Jumlah Transaksi
Shift 1	0.2587	
Shift 2	0.2864	

**Table 11.** Hasil *Pattern Similarity* terhadap Jumlah Transaksi

Data	<i>Pattern Similarity</i>	Jumlah Transaksi (%)
Shift 1		78.11
Shift 2		83.12

**Table 12.** Hasil *Pattern Similarity* Terhadap Total Pendapatan

Data	<i>Pattern Similarity</i>	Jumlah Transaksi (%)
Shift 1		76.29
Shift 2		85.79

yang diperoleh termasuk kedalam interpretasi prediksi yang masuk akal (reasonable). Nilai MAPE pada model *shift 1* lebih kecil dari *shift 2*, menunjukkan bahwa model *shift 1* lebih baik dalam memprediksi dibandingkan dengan model *shift 2*.

#### Pengujian dengan *Pattern Similarity*

*Pattern Similarity* menunjukkan seberapa cocok pola data antara prediksi dengan sebenarnya terlepas dari besar *error*-nya. Hasil pengujian *Pattern Similarity* terletak pada tabel 11 & 12 ini.

Terlepas dari besarnya *error* yang dihasilkan, kedua model dapat memprediksi dengan pola mirip dengan sebenarnya.

#### Kesimpulan

Model XGBoost dalam penelitian ini dapat digunakan untuk melakukan prediksi terhadap jumlah transaksi dan total pendapatan. Hasil evaluasi menunjukkan bahwa model memiliki performa yang baik, dengan nilai  $R^2$  yang menunjukkan tingkat kecocokan yang cukup memadai antara hasil prediksi dan data sebenarnya. RMSE dan MAPE yang diperoleh menunjukkan bahwa model *shift 1* memiliki tingkat

*error* yang lebih rendah dibandingkan dengan model *shift 2*, menandakan bahwa model *shift 1* ini lebih baik dalam memprediksi jumlah transaksi dan total pendapatan. Selain itu, *Pattern Similarity* menunjukkan bahwa pola prediksi yang dihasilkan oleh model memiliki tingkat kesamaan yang tinggi dengan data sebenarnya, meskipun terdapat perbedaan dalam besarnya *error*. Secara keseluruhan, model XGBoost yang dioptimalkan dengan *Bayesian Optimization* memberikan hasil yang memuaskan dan dapat diandalkan untuk keperluan manajemen operasional jaringan restoran CV Balibul.

#### Daftar Pustaka

- Chen T, Guestrin C. XGBoost. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM; 2016. p. 785-94.
- Ahmed S, et al. The Deep Learning ResNet101 and Ensemble XGBoost Algorithm with Hyperparameters Optimization Accurately Predict the Lung Cancer. Applied Artificial Intelligence. 2023 December;37(1).
- Gupta P, Bagchi A. Introduction to Pandas; 2024. p. 161-96.
- Shi R, Xu X, Li J, Li Y. Prediction and Analysis of Train Arrival Delay Based on XGBoost and Bayesian Optimization. Applied Soft Computing. 2021 September;109:107538.
- Ledmaoui Y, Maghraoui AE, Aroussi ME, Saadane R, Chebak A, Chehri A. Forecasting Solar Energy Production: A Comparative Study of Machine Learning Algorithms. Energy Reports. 2023 November;10:1004-12.
- Chicco D, Warrens MJ, Jurman G. The Coefficient of Determination R-squared is More Informative than SMAPE, MAE, MAPE, MSE and RMSE in Regression Analysis Evaluation. PeerJ Computer Science. 2021 July;7:e623.
- Vivas E, Allende-Cid H, Salas R. A Systematic Review of Statistical and Machine Learning Methods for Electrical Power Forecasting with Reported MAPE Score. Entropy. 2020 December;22(12):1412.
- Sheugh L, Alizadeh SH. A Note on Pearson Correlation Coefficient as a Metric of Similarity in Recommender System. In: 2015 AI & Robotics (IRANOPEN). IEEE; 2015. p. 1-6.