Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1649−1661
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8657

Implementasi dan Evaluasi Model Machine Learning untuk Optimalisasi Prediksi Penjualan Produk Kue Kering

Muhammad Abror Auliya Hilmi*, Ajib Susanto

Fakultas Ilmu Komputer, Teknik Informatika, Universitas Dian Nuswantoro, Semarang, Indonesia
Email: [1,*]111202214677@mhs.dinus.ac.id, [2,]ajib.susanto@dsn.dinus.ac,id
Email Penulis Korespondensi: dedykurniawan@unsri.ac.id
Submitted: **07/11/2025**; Accepted: **06/12/2025**; Published**: 08/12/2025**

Abstrak−Sektor ritel modern, seperti Transmart, menghadapi kesulitan dalam mempertahankan kestabilan penjualan karena perubahan perilaku konsumen, variasi produk, serta karakteristik lokasi yang berbeda. Untuk mengatasi isu ini, penelitian ini menyarankan penggunaan algoritme machine learning Extreme Gradient Boosting (XGBoost) untuk memprediksi volume penjualan produk ritel berdasar data historis tahun 2024–2025. Penelitian ini menggunakan kerangka kerja CRISP-DM yang mencakup langkah-langkah seperti Pemahaman Bisnis, Pemahaman Data, Persiapan Data, Pemodelan, Evaluasi, dan Implementasi. Proses pembersihan dan pra-pemrosesan data dilakukan melalui langkah-langkah seperti pembersihan data, pengkodean label, pemilihan fitur, serta pembagian data dengan rasio 80:20. Model selanjutnya dievaluasi menggunakan metrik Mean Absolute Error (MAE) dan koefisien determinasi (R^2) untuk menilai tingkat akurasi prediksi. Temuan penelitian menunjukkan bahwa XGBoost mampu memahami pola penjualan dengan baik dan menghasilkan prediksi yang tepat guna mendukung strategi pengambilan keputusan di sektor ritel, khususnya dalam hal perencanaan stok dan optimalisasi penjualan. Oleh karena itu, penerapan pendekatan prediktif berbasis data ini diharapkan dapat membantu perusahaan dalam meningkatkan manajemen operasional serta daya saing di pasar.

Kata Kunci: Prediksi Penjualan; XGBoost; Machine Learning; CRISP-DM; Ritel

Abstract−The modern retail sector, such as Transmart, faces difficulties in maintaining stable sales performance due to changes in consumer behavior, variations in product types, and differing store characteristics. To address this issue, this study proposes the use of the Extreme Gradient Boosting (XGBoost) machine learning algorithm to predict retail product sales volumes based on historical data from 2024–2025. The research utilizes the CRISP-DM framework, which consists of the following stages: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. The data cleaning and preprocessing processes involve several steps such as data cleaning, label encoding, feature selection, and data splitting with an 80:20 ratio. The model is further evaluated using the Mean Absolute Error (MAE) and the coefficient of determination (R^2) metrics to assess prediction accuracy. The findings indicate that XGBoost is capable of effectively capturing sales patterns and generating accurate predictions to support decision-making strategies in the retail sector, particularly in stock planning and sales optimization. Therefore, the implementation of this data-driven predictive approach is expected to assist companies in enhancing operational management as well as improving competitiveness in the market.

Keywords: Prediction Market; XGBoost; Machine Learning; CRISP-DM; Retail

1. PENDAHULUAN

Teknologi Informasi adalah suatu teknologi yang digunakan untuk mengolah data, termasuk memproses, mendapatkan,menyusun, menyimpan, memanipulasi data dalam berbagai cara untuk menghasilkan informasi yang berkualitas, yaitu informasi yang relevan, akurat dan tepat waktu, yang digunakan untuk keperluan pribadi, bisnis, dan pemerintahan dan merupakan informasi yang strategis untuk pengambilan Keputusan [1]. Kondisi saat ini bila dinilai dari faktor pendorong terjadinya transformasi digital, dapat dikatergorikan dalam kategori faktor pertama, perubahan regulasi. Munculnya pandemi covid-19 menyebabkan pemerintah mengeluarkan regulasi baru bahwa selama masa pandemi semua dikerjakan melalui media digital / dalam jaringan sehingga mau tidak mau semua harus mengikuti regulasi tersebut [2]. Salah satu entitas ritel nasional yang menghadapi tantangan besar dalam era ini adalah Transmart, jaringan ritel yang beroperasi di berbagai kota dengan karakteristik pasar yang berbeda. Permasalahan utama yang dihadapi Transmart terletak pada ketidakstabilan volume penjualan, yang dipengaruhi oleh berbagai faktor seperti waktu, lokasi gerai, harga, variasi produk, hingga preferensi konsumen yang dinamis. Oleh karena itu, analisis data penjualan yang akurat menjadi kebutuhan mendesak untuk mendukung perencanaan stok, penyusunan strategi promosi, dan pengambilan keputusan manajerial yang tepat sasaran.

Dalam beberapa tahun terakhir, perkembangan teknologi *machine learning* membuka peluang baru dalam implementasi algoritma regresi berbasis data [3]. Penelitian ini berupaya menjawab permasalahan tersebut melalui penerapan model XGBoost (Extreme Gradient Boosting) sebagai pendekatan utama dalam memprediksi penjualan produk ritel Transmart. *XGBoost (Extreme Gradient Boosting)* merupakan salah satu Algoritma *Machine Learning* yang biasa digunakan untuk prediksi [4]. XGBoost (Extreme Gradient Boosting) adalah pengembangan lebih lanjut dari algoritma Gradient Tree Boosting berbasis ensemble, yang dapat secara efektif menangani masalah machine learning dengan skala yang lebih besar [5]. Model ini dipilih karena kemampuannya dalam menangani data kompleks yang melibatkan banyak variabel serta menghasilkan akurasi tinggi pada proses prediksi. Model yang dihasilkan diharapkan dapat memberikan prediksi yang akurat sekaligus transparan [6]. Kebaruan utama penelitian ini terletak pada penerapan model XGBoost secara menyeluruh terhadap data riil multi-lokasi dan multi-produk, yang belum banyak dilakukan pada konteks ritel nasional di Indonesia. Algoritma Machine Learning (ML), khususnya Extreme

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1649−1661
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8657

Gradient Boosting (XGBoost), memiliki kapabilitas tinggi dalam melakukan prediksi akurat melalui pendekatan pemodelan non-linear dan efisiensi dalam pengolahan data [7].

Data mining adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu [8]. kebutuhan akan model machine learning menjadi semakin mendesak untuk menghasilkan prediksi yang lebih akurat, adaptif, dan relevan dalam konteks dunia nyata[3]. Proses ini memanfaatkan teknik statistik, matematika, dan pembelajaran mesin (*machine learning*) untuk mengekstraksi pengetahuan yang sebelumnya tidak diketahui dan tidak dapat diperoleh secara manual [9]. Dalam penelitian ini, data penjualan yang digunakan telah melalui proses analisis eksploratif (Exploratory Data Analysis/EDA) untuk memahami pola umum dan mendeteksi anomali.Exploratory Data Analysis (EDA) adalah proses menganalisis dan menampilkan data bertujuan mendapatkan pemahaman yang lebih baik tentang wawasan dari data [10].

CRISP-DM (Cross Industry Standard Process for Data Mining) suatu standarisasi pemrosesan data mining yang telah dikembangkan dimana data yang ada akan melewati setiap fase terstruktur dan terdefinisi dengan jelas dan efisien. Selain menerapkan suatu model dalam proses penambangan data, pemilihan algoritma sangat mempengaruhi terhadap komparasi kinerja metode data mining [11]. Oleh karena itu, diperlukan strategi pemilihan algoritma yang tepat sesuai karakteristik data dan tujuan analisis, sehingga model prediktif yang diterapkan dapat memberikan hasil yang akurat, efisien, serta memiliki kemampuan generalisasi yang baik terhadap data baru. Implementasi CRISP-DM pada penelitian ini berperan sebagai pedoman untuk memastikan setiap proses pemodelan dilakukan secara terarah mulai dari pemahaman bisnis, pengolahan data, pemodelan, hingga proses evaluasi dan penerapan hasil akhir.

Penelitian terdahulu menunjukkan bahwa algoritma machine learning memiliki kemampuan prediktif yang kuat di berbagai domain. Muhammad Fakhrul Reza dan Ghufron dalam jurnal "Implementasi Model Gated Recurrent Unit (GRU) atau Extreme Gradient Boosting (XGBoost) untuk Prediksi Harga Cryptocurrency Ethereum" membuktikan bahwa GRU dan XGBoost mampu memprediksi harga Ethereum dengan kesalahan rendah, dengan GRU sebagai model terbaik. Muhamad Amhar Rayadin dkk. melalui "Implementasi Ensemble Learning Metode XGBoost dan Random Forest untuk Prediksi Waktu Penggantian Baterai Aki" [4] menunjukkan bahwa ensemble XGBoost–Random Forest meningkatkan akurasi prediksi dibandingkan model tunggal. Adapun Fatika Lovina Febrianti dkk. dalam "Implementasi Metode XGBoost dan SHAP untuk Klasifikasi dan Analisis Faktor Risiko Penyakit Diabetes Mellitus"[6] menemukan bahwa XGBoost menghasilkan akurasi klasifikasi yang sangat tinggi sekaligus mampu mengungkap faktor risiko utama melalui SHAP. Temuan tersebut menegaskan bahwa XGBoost dan turunannya relevan untuk diadopsi dalam penelitian ini guna mendukung prediksi berbasis data.

Kebaruan lain yang menjadi pembeda penelitian ini adalah adanya perbandingan hasil prediksi antara model XGBoost dengan metode CAGR (Compound Annual Growth Rate). Compound Annual Growth Rate(CAGR) merupakan alat bantu kuantitatif yang dapat digunakan dalam konteks tren analisis pasar [12]. Analisi tren digunakan untuk mengetahui kecenderungan perkembangan laba dari tahun ke tahun, volatilitas digunakan untuk mengukur tingkat kestabilan, CAGR untuk menghitung pertumbuhan rata-rata tahunan, dan metode least squares untuk meramalkan laba di masa mendatang [13]. Metode CAGR digunakan untuk menghitung laju pertumbuhan tahunan rata-rata penjualan. Hasil perhitungan menunjukkan bahwa berdasarkan CAGR, penjualan Transmart mengalami pertumbuhan konstan sebesar 52,41 persen per tahun. Namun, hasil yang diperoleh dari model XGBoost memperlihatkan nilai berbeda signifikan, yakni -0,17 persen, yang menandakan adanya potensi penurunan penjualan di tahun mendatang. Perbedaan hasil antara dua pendekatan ini menjadi temuan penting dalam penelitian, karena mengindikasikan bahwa pertumbuhan penjualan tidak selalu bersifat linier dan dapat dipengaruhi oleh berbagai faktor eksternal seperti perilaku konsumen, tren musiman, serta kondisi ekonomi lokal di setiap cabang.

Analisis tren pasar menjadi penting dalam menginformasikan keputusan bisnis yang tepat waktu dan efektif. Selain itu, semakin banyak bisnis yang mengandalkan data untuk mendukung pengambilan keputusan mereka, sehingga memperkuat pentingnya memiliki strategi berbasis data yang kuat [14]. Dari sisi implementasi praktis, penelitian ini memberikan kontribusi signifikan bagi sektor ritel Indonesia, khususnya dalam konteks pengambilan keputusan berbasis data (data-driven decision making). Dengan hasil prediksi yang akurat, pihak manajemen Transmart dapat memanfaatkan model ini untuk memperkirakan permintaan pasar secara lebih tepat, mengoptimalkan distribusi produk antar cabang, serta menyesuaikan strategi penetapan harga dan promosi. Model XGBoost juga dapat diintegrasikan dengan sistem Enterprise Resource Planning (ERP) yang sudah ada untuk mendukung pengambilan keputusan secara otomatis dan real-time.

2. METODOLOGI PENELITIAN

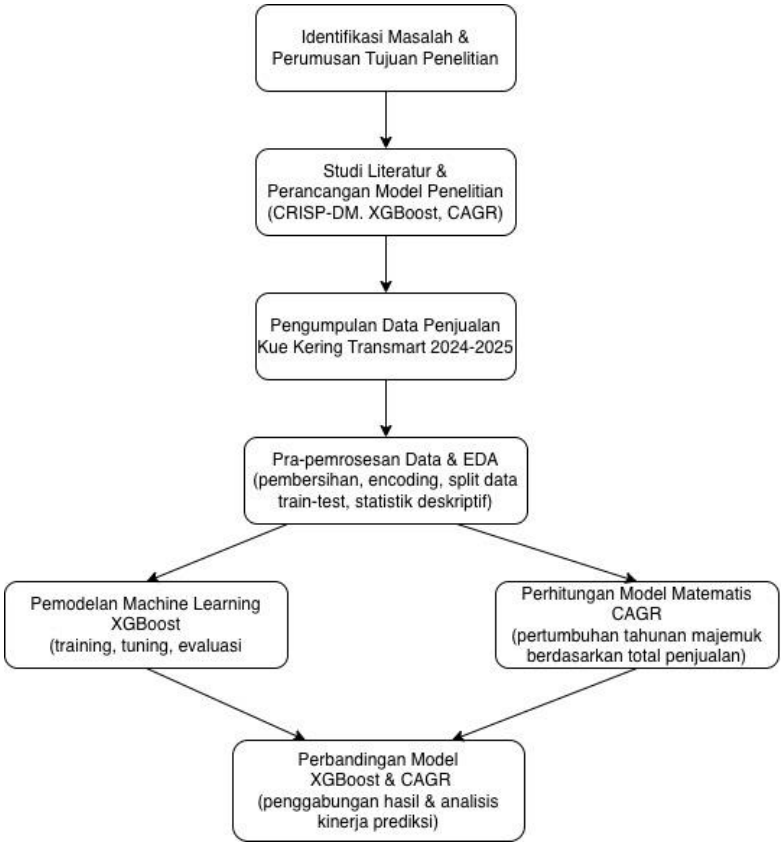
2.1 Tahapan Penelitian

2.1.1 Alur Tahapan Penelitian

Alur penelitian menguraikan proses langkah yang diterapkan secara terstruktur, dimulai dari pengenalan masalah dan penetapan tujuan, penelitian pustaka, pengumpulan serta pemrosesan awal data penjualan kue kering Transmart, sampai pada tahap pemodelan dengan XGBoost dan penghitungan CAGR, yang kemudian dievaluasi performanya. Rincian menyeluruh mengenai penelitian tersebut disajikan dalam bentuk diagram pada Gambar 1, yang

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1649−1661
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8657

memungkinkan pembaca untuk mengamati urutan proses penelitian dan bagaimana setiap tahap saling berhubungan untuk menghasilkan model prediksi penjualan yang tepat.



Gambar 1. Diagram Alur Penelitian

Gambar 1 menunjukkan diagram alir yang menggambarkan tahapan penelitian dimulai dari mengenali masalah ketidakpastian yang berkaitan dengan penjualan kue kering di Transmart serta menyiapkan tujuan dari penelitian ini. Proses ini dilanjutkan dengan tinjauan literatur dan pengembangan model dengan menggunakan kerangka kerja CRISP-DM, di mana fokus utama dari penelitian ialah membangun dan menyempurnakan model machine learning XGBoost sebagai alat untuk memprediksi penjualan, sementara metode matematis CAGR digunakan sebagai model referensi. Selanjutnya, data penjualan kue kering untuk tahun 2024–2025 berhasil dikumpulkan dan mengalami proses tahapan pra-pemrosesan dan analisis data eksploratif, mencakup pembersihan, pengkodean, pembagian data menjadi set pelatihan dan pengujian, serta melakukan analisis statistik deskriptif untuk mempersiapkan data yang sesuai dengan algoritma XGBoost. Setelah itu, penelitian diarahkan pada pemodelan machine learning menggunakan XGBoost yang melibatkan proses pelatihan, penyesuaian hiperparameter, dan penilaian kinerja, kemudian dibandingkan dengan hasil dari perhitungan CAGR. Pada tahap akhir adalah penilaian dan perbandingan hasil dari kedua model tersebut untuk mengukur sejauh mana XGBoost efektif dalam memprediksi penjualan kue kering serta menentukan model yang paling tepat digunakan untuk proyeksi di masa mendatang.

2.1.2 Implementasi Cross-Industry Standard Process for Data Mining (CRISP-DM) terhadap performa model XGBoost.

Kerangka kerja CRISP-DM menjelaskan proses siklus pemrosesan data secara berulang, dimulai dari pemahaman bisnis, pemahaman data, persiapan data, pemodelan, evaluasi, hingga penempatan yang semuanya terfokus pada data. Dalam studi ini, CRISP-DM diaplikasikan untuk merancang proses pembuatan model prediksi penjualan dengan XGBoost, sehingga setiap Langkah dari menentukan kebutuhan bisnis hingga menilai hasil prediksi dilaksanakan dengan cara yang sistematis dan terarah. Penerapan utama CRISP-DM diutamakan pada fase pemodelan dan evaluasi, di mana algoritma XGBoost dibangun, disesuaikan, dan diukur kinerjanya menggunakan metrik kesalahan untuk menilai seberapa baik model dapat memprediksi penjualan kue kering.

Cross-Industry Standard Process for Data Mining (CRISP-DM) merupakan kerangka kerja yang sering diterapkan dalam proyek data sains dan penambangan data. CRISP-DM terdiri atas enam fase sebagai berikut:

a. Bussines Understanding, mengidentifikasi masalah dan tujuan bisnis.
b. Data Understanding, mengeksplorasi karakteristik data penjualan yang akan digunakan sebagai input model XGBoost.
c. Data Preparation, melakukan pembersihan, transformasi, dan pengolahan fitur agar sesuai dengan kebutuhan algoritma XGBoost.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1649−1661
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8657

d. Modeliing, menciptakan dan mengkonfigurasi model XGBoost, termasuk penyesuaian hiperparameter dan metode pelatihan.
e. Evaluation, menilai kemampuan model XGBoost dengan menggunakan metrik seperti kesalahan prediksi dan koefisien determinasi untuk memastikan model mencapai tujuan bisnis.
f. Deployment, menyiapkan hasil dari model prediksi XGBoost untuk digunakan sebagai dasar dalam pengambilan keputusan mengenai perencanaan stok, distribusi, dan strategi promosi.

2.1.3 Bussines Understanding

Fase Business Understanding terdiri dari dua proses yaitu penentuan tujuan bisnis dan pendefinisian tujuan proyek.[15] Fase Busiines Understading dilakukan melalui wawancara terhadap pemilik usaha kue kering. Tujuan bisnis dari pemilik usaha adalah untuk mengetahui prediksi penjualan produk tahun mendatang, yang mana dimulai dari prediksi tahun 2026 dan diharapkan mampu memprediksi hingga tahun tahun selanjutnya. Tujuan proyek ini adalah untuk memprediksi hasil penjualan kue kering dari dataset penjualan produk tahun 2024 dan 2025 untuk prediksi produksi penjualan tahun 2026 dan seterusnya, tetapi pada penelitian ini difokuskan untuk prediksi produksi dan penjualan tahun 2026.

2.1.4 Data Understanding

Fase Data Understanding tahap pertama terdiri dari empat proses yaitu pengumpulan data, pengintegrasian data, pendeskripsian data, dan pengeksplorasian data. Pengumpulan data dilakukan melalui studi dokumen.[15] Dokumen yang digunakan dalam penelitian ini meliputi catatan penjualan produk kue kering yang terjual di gerai Transmart dari tahun 2024 hingga tahun 2025.

2.1.5 Data Preparation

Fase Data Preparation tahap pertama terdiri dari empat proses yaitu pemformatan data, pembersihan data, pengelompokan data, dan pemilihan data. Pembersihan data dilakukan untuk mengatasi nilai yang hilang dan mengatasi pencilan. Pengelompokan data dilakukan untuk mengelompokkan data penjualan berdasarkan produk dan tanggal penjualan [15]Pemilihan data dilakukan untuk memilih data dari produk yang akan diprediksi. Produk yang akan diprediksi sebanyak 838 total penjualan produk, dengan 11 total varian produk.

Pada tahap ini dilakukan serangkaian proses transformasi dan pembersihan data agar dataset siap digunakan dalam pemodelan machine learning menggunakan algoritma XGBoost. Tahapan pra-pemrosesan data merupakan fase penting dalam kerangka kerja CRISP-DM karena menentukan kualitas dan akurasi model yang dihasilkan. Adapun langkah-langkah yang dilakukan adalah sebagai berikut.

a. Exploratory Data Analysis (EDA)
   EDA dilakukan untuk memperoleh pemahaman awal terhadap struktur, pola, serta anomali yang terdapat dalam datapenjualan Transmart. Analisis ini dilakukan melalui pemeriksaan nilai minimum, maksimum, rata-rata, dan sebaran data (descriptive statistics), serta visualisasi berupa grafik batang, diagram sebar (scatter plot), dan heatmap. Tahapan ini bertujuan untuk mengidentifikasi tren penjualan per tahun, persebaran nilai harga dan jumlah barang terjual, serta mendeteksi adanya outlier atau nilai ekstrem yang dapat memengaruhi hasil pemodelan. Selain itu, EDA juga digunakan untuk menilai keterkaitan antar-fitur seperti hubungan antara harga unit dengan total penjualan, serta perbedaan pola penjualan antar-lokasi. Hasil dari tahap ini menjadi dasar dalam menentukan strategi pembersihan dan pemilihan fitur yang relevan bagi model XGBoost.
b. Pembersihan Data (Data Cleaning)
   Langkah pertama dilakukan dengan menghapus nilai kosong (*missing values*) serta duplikasi data menggunakan fungsi dropna() dan drop_duplicates(). Tindakan ini bertujuan untuk menjaga integritas dataset dan mencegah bias model akibat adanya entri yang tidak lengkap atau berulang.
c. Transformasi Data Kategorikal (Label Encoding)
   Kolom dengan tipe data kategorikal, seperti lokasi dan nama barang, diubah menjadi representasi numerik menggunakan teknik Label Encoding melalui pustaka LabelEncoder. Transformasi ini diperlukan karena algoritma XGBoost hanya dapat memproses data numerik. Proses pengkodean dilakukan secara konsisten antara data pelatihan dan data prediksi tahun 2026 untuk memastikan keseragaman representasi kategori.
d. Penetapan Variabel Target
   Variabel yang menjadi target prediksi adalah jumlah barang terjual (total qty). Kolom tersebut dikonversi menjadi tipe data float agar dapat diproses secara numerik oleh model. Penentuan variabel target dilakukan berdasarkan tujuan penelitian, yaitu memprediksi jumlah penjualan produk.
e. Seleksi Fitur dan Pemisahan Data
   Setelah proses pembersihan dan pengkodean selesai, dilakukan pemilihan fitur yang relevan terhadap target, meliputi variabel tahun, harga unit, lokasi enc, dan barang enc. Selanjutnya, dataset dibagi menjadi dua bagian, yakni data latih (training data) sebesar 80 persen dan data uji (testing data) sebesar 20 persen menggunakan fungsi train_test_split. Pembagian ini dimaksudkan agar model dapat dievaluasi secara objektif terhadap data yang belum pernah dilihat sebelumnya.
f. Normalisasi dan Konsistensi Tipe Data

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1649−1661
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8657

Seluruh kolom numerik diperiksa untuk memastikan keseragaman tipe data, khususnya pada atribut harga unit dan total harga. Normalisasi dilakukan guna menghindari ketidaksesuaian format desimal atau pemisah ribuan yang dapat memengaruhi hasil perhitungan model.

g. Penerapan Encoder pada Dataset Baru
Saat model diaplikasikan pada data prediksi tahun 2026, dilakukan penerapan kembali encoder yang telah dilatih sebelumnya agar nilai kategori pada data baru memiliki makna yang sama dengan data pelatihan. Langkah ini memastikan hasil prediksi tetap konsisten dan dapat dibandingkan secara valid antar tahun.

2.1.6 Modelling

Fase Modeling dilakukan secara terpisah untuk setiap produk yang dipilih. Fase Modelingterdiri dari proses-prosesseperti pembagian data menjadi data latih dan data uji, pelatihan model XGBoost menggunakan data latih, peramalan terhadap data uji, dan perhitungan nilai evaluasi performa model. Fase Modeling dilakukan melalui tahap pengujian pelatihan model [15].

2.1.7 Evaluation

Fase Evaluation yaitu evaluasi hasil pelatihan model XGBoost [15]. Tahap evaluasi merupakan proses untuk menilai kinerja model machine learning yang telah dibangun pada fase modeling. Evaluasi dilakukan untuk mengetahui sejauh mana model mampu memprediksi nilai target secara akurat dan konsisten terhadap data uji, serta memastikan bahwa model tidak mengalami overfitting atau underfitting. Pada penelitian ini, proses evaluasi dilakukan menggunakan dua metrik utama, yaitu Mean Absolute Error (MAE) dan Koefisien Determinasi (R^2). Metrik MAE digunakan untuk mengukur rata-rata kesalahan absolut antara nilai aktual dengan nilai hasil prediksi, sedangkan R^2 digunakan untuk mengetahui seberapa besar variasi data aktual yang dapat dijelaskan oleh model.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (1)$$

Pada persamaan (1), n merupakan jumlah observasi pada data uji, y_i merupakan nilai actual (ground truth), $\hat{y}_i$ merupakan nilai hasil prediksi model.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (2)$$

Pada persamaan (2), y merupakan rata-rata nilai aktual di data uji, $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ merupkan "residual sum of squares"/ jumlah kudrat galat, $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ merupakan "total sum of squares"/ variasi total terhadap rataan.

2.1.8 Deployment

Fase Deployment terdiri dari proses melakukan peramalan untuk masa yang akan datang menggunakan model terpilih.[15]. Tahap deployment merupakan fase akhir dari proses CRISP-DM, di mana model yang telah dievaluasi dan dinyatakan memiliki performa terbaik diimplementasikan pada data baru untuk menghasilkan prediksi aktual. Pada penelitian ini, model XGBoost yang telah dilatih dan divalidasi diterapkan pada dataset tahun 2026 yang telah melalui tahap pra-pemrosesan dengan prosedur pengkodean (encoding) yang sama seperti data pelatihan.

2.1.9 CAGR (*Compound Annual Growth Rate*)

Selain menggunakan algoritma machine learning XGBoost, penelitian ini juga menerapkan metode matematis Compound Annual Growth Rate (CAGR) sebagai alat bantu pembanding hasil prediksi penjualan. CAGR merupakan salah satu pendekatan kuantitatif yang digunakan untuk mengukur tingkat pertumbuhan rata-rata tahunan suatu variabel dalam periode tertentu, dengan asumsi bahwa pertumbuhan terjadi secara konstan setiap tahun.

$$CAGR = \left(\frac{V_t}{V_0}\right)^{\frac{1}{t}} - 1 \quad (3)$$

Pada Persamaan (3) menunjukkan perhitungan CAGR, di mana V_t adalah nilai akhir pada periode pengamatan terkahir, yaitu 2025 V_0 adalah nilai awal periode, yaitu 2024 t adalah jumlah tahun atau periode waktu, yaitu 2

3. HASIL DAN PEMBAHASAN

3.1 Dataset

dataset penelitian ini merekam transaksi ritel kue kering Transmart pada horizon 13 Januari 2024 hingga 30 April 2025. Struktur workbook memuat dua lembar data, yakni Data Penjualan (sebagai fact table utama untuk pemodelan) dan Referensi Barang (sebagai tabel dimensi/metadata produk). Lembar Data Penjualan terdiri atas 837 baris observasi dengan cakupan spasial 18 gerai (cabang Transmart) dan 12 varian produk, sehingga secara informatik memberikan representasi yang memadai untuk analisis musiman, heterogenitas lokasi, serta diferensiasi produk. Berikut adalah 5 tabel teratas pada tahun 2024 dan 2025.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1649−1661
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8657

Tabel 1. Sample Dataset 2024

Tgl order	Lokasi	Nama barang	Total qty	Harga unit	Total Harga
2024-04-15	Ambarukmo plaza	Varian 1	1	65000	65000
2024-04-15	Ambarukmo plaza	Choco chocolate	1	71000	71000
2024-04-15	Ambarukmo plaza	Varian 4	2	65000	130000
2024-04-15	Ambarukmo plaza	Varian 3	2	65000	130000
2024-04-15	Ambarumo plaza	Varian 2	7	65000	455000

Tabel 1, menampilkan rekaman transaksi berbutir (atomic transaction records) yang digunakan sebagai fact table pada pipeline pemodelan. Setiap baris mengandung atribut waktu (Tgl order, bertipe datetime yang menjadi sumber rekayasa fitur bulan/kuartal/dow), dua dimensi kategorikal (Lokasi dan Nama barang) yang selanjutnya direpresentasikan secara numerik melalui label-encoding persisten untuk menghindari category drift, serta tiga métrik numerik (Total qty, Harga unit, Total Harga) yang telah dikonversi ke tipe numerik bersih (pembersihan pemisah ribuan dan simbol mata uang). Secara semantik, baris-baris contoh memperlihatkan konsistensi aritmatika yang berfungsi sebagai aturan integritas domain pada tahap data validation (ETL).

Tabel 2. Sample Dataset 2025

Tgl Order	Lokasi	Nama Barang	Total qty	Harga Unit	Total Harga
2025-03-04	Transmart Telogorejo	Nastar Keju	2	52000	104000
2025-01-03	Transmart Gresik	Varian 1	1	72000	72000
2025-01-03	Transmart Gresik	Nastar Keju	1	52000	52000
2025-01-03	Transmart Gresik	Nastar	1	52000	52000
2025-01-03	Transmart Gresik	Putri Salju	1	45000	45000

Tabel 2, menyajikan sampel transaksi pada horizon tahun 2025 dengan skema identik, sehingga isomorfik terhadap struktur 2024 dan kompatibel untuk penggabungan (union) maupun split train–test yang reprodusibel. Keberadaan atribut waktu dengan rentang tahun berbeda memfasilitasi validasi temporal (mis. latih pada awal periode, uji pada akhir periode) dan penyusunan fitur musiman yang konsisten antar-tahun; sementara kombinasi Lokasi dan Nama barang menyediakan sumbu segmentasi untuk slice-based validation (menilai kinerja model pada segmen berkontribusi tinggi maupun ekor). Seperti pada 2024, contoh baris memperlihatkan invarian bisnis yang menjadi in-line check selama ETL dan sumber fitur intensitas harga(dengan proteksi pembagi nol) pada tahap Feature Engineering.

3.2 EDA (Exploratory Data Analysis)

3.2.1 Data Dictionary

Tabel 3. Data Dictionary

Fitur	Tipe Interferensi	Misiing_%	Unik	Contoh Nilai
Tgl Order	Datetime64[ns]	0.0	52	2024-04-15
Lokasi	Object	0.0	18	Ambarumo Plaza
Nama Barang	Object	0.0	12	Varian 1
Total Qty	Float 64	0.0	17	1.0
Harga Satuan	Float 64	0.0	10	65000
Total Harga	Float 64	0.0	85	113000
Tahun	Int 64	0.0	2	2025

Tabel 3, memetakan skema dataset: tipe inferensi setiap kolom, proporsi nilai hilang (missing%), jumlah nilai unik, dan contoh nilai. Secara metodologis, data dictionary memastikan validitas skema sebelum pemodelan (mencegah type drift dan inflasi kategori saat encoding). Pada data ini kolom inti tidak memiliki nilai hilang sehingga tidak memerlukan imputasi pada tahap awal.

3.2.2 Ringkasan Statistik Numerik (qty, harga, nominal)

Tabel 4. Ringkasan Statistik Numerik (qty, harga, nominal)

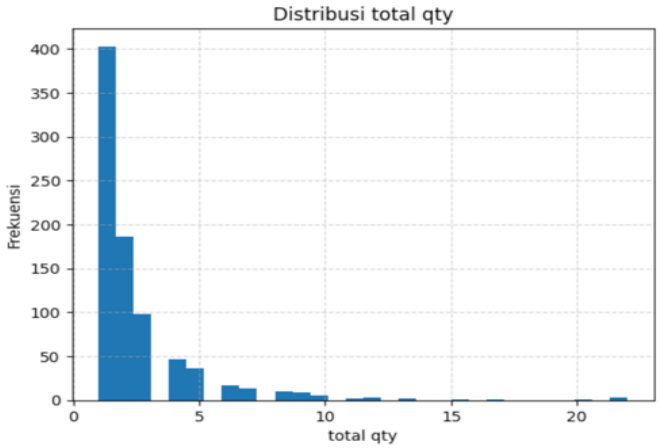
Fitur	Count	Mean	Std	Min	25%	50%	75%	Max	Skewness	Missing_%
Total Qty	837.0	2.440860	2.491690	1.0	1.0	2.0	3.0	22.0	3.654588	0.0
Harga	837.0	61548.387097	9831.674859	45000.0	52000.0	65000.0	72000.0	76000.0	0.248677	0.0
Total	837.0	150881.720430	153429.282100	45000.0	63000.0	94000.0	180000.0	1430000.0	3.220069	0.0

Pada Tabel 4, berbasis n = 837 observasi pada tiga fitur inti total qty, harga satuan, dan total harga menunjukkan 0% nilai hilang, sehingga kualitas data memadai untuk analisis lanjutan dan pemodelan. Keberadaan metrik pemusatan (mean/median), sebaran (simpangan baku, kuartil), serta shape distribusi (skewness)

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1649−1661
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8657

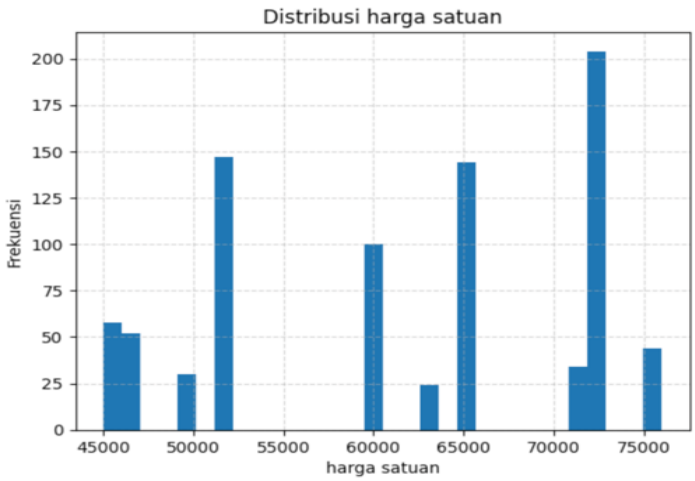
memungkinkan karakterisasi statistik yang komprehensif sesuai kaidah EDA dalam informatika. Fitur total qty memiliki rata-rata 2,44 (median 3; Q1–Q3; 1–3; maksimum 22) dengan skewness 3,65, menandakan distribusi sangat menceng ke kanan: mayoritas transaksi membeli 1–3 unit, sementara sebagian kecil pembelian dalam jumlah besar membentuk ekor kanan panjang. Dengan IQR=2, ambang upper fence berbasis IQR = 6 unit, sehingga transaksi dengan kuantitas >6 layak ditandai sebagai kandidat outlier operasional. Fitur harga satuan relatif stabil (mean = 61,5K; median 65K; Q1-Q3: 52K–72K; skewness −0,25), mengindikasikan pita harga yang sempit dan mendekati simetris. Stabilitas ini penting karena mengurangi variansi yang tak perlu pada pembentukan target total harga, serta meminimalkan kebutuhan imputasi atau normalisasi agresif. Fitur total harga—hasil interaksi kuantitas dan harga memperlihatkan rata-rata = 150,9K (median 140K; Q1–Q3: 94K–180K; maksimum 1,43M) dengan skewness 3,22, menunjukkan ekor kanan panjang yang wajar akibat transaksi ber-qty tinggi. IQR=86K memberi upper fence = 309K; nilai di atas ambang ini lebih tepat diperlakukan sebagai transaksi bernilai tinggi ketimbang kesalahan, namun tetap perlu flagging untuk kontrol pengaruh terhadap model.

3.2.3 Distribusi Variabel Numerik



Gambar 2. Distribusi Total Qty

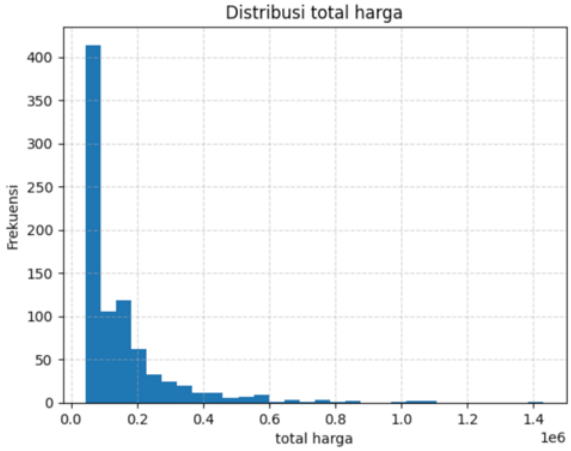
Pada Gambar 2, Distribusi menunjukkan karakter right-skewed yang kuat dengan modus pada 1–3 unit; frekuensi turun drastis setelah nilai 3 dan membentuk ekor kanan panjang hingga >20 unit. Sifat diskret dan zero-truncated (nilai minimum 1) memunculkan puncak-puncak pada bilangan bulat. Pola ini mengindikasikan pembelian ritel tipikal berskala kecil dengan sejumlah kecil transaksi borongan berperan sebagai pengamatan berpengaruh (high leverage). Secara diagnostik, ambang kandidat outlier berbasis IQR berada sekitar qty >-6; observasi di atas nilai ini sebaiknya ditandai untuk capping/winsorization atau dijadikan fitur indikator (mis. bulk_flag = qty>=7).



Gambar 3. Distribusi Harga Satuan

Pada Gambar 3, Sebaran tampak sempit dan relatif stabil, dengan multi-modal ringan pada beberapa titik harga (45k, 52k, 60k, 65k, 72–75k). Puncak-puncak tersebut lazim pada kebijakan penetapan harga bertangga/promosi, sehingga distribusi mendekati simetris (sedikit left-skew) tanpa indikasi anomali ekstrem. Karena varians rendah dan tidak ada nilai hilang, fitur ini siap pakai sebagai numerik kontinu; alternatifnya, untuk menangkap struktur tangga harga, dapat dibentuk price-tier encoding (mis. one-hot untuk kelas harga dominan atau ordinal binning per 5.000/10.000).

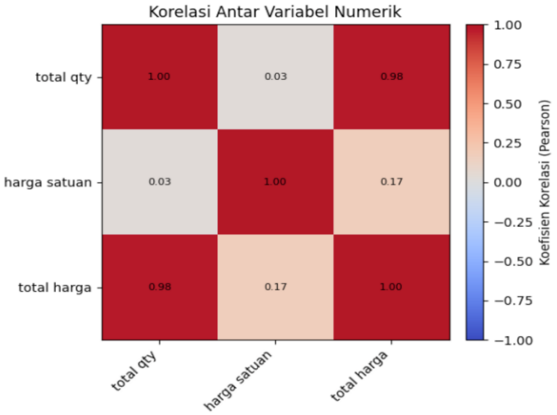
Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1649−1661
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8657



Gambar 4. Distribusi Total Harga

Pada Gambar 4, Distribusi sangat right-skewed: massa utama berada pada rentang 90k–180k, namun ekor kanan meluas hingga >1 juta. Bentuk ini konsisten dengan sifat perkalian total qty × harga satuan; kombinasi kuantitas besar pada pita harga relatif konstan menghasilkan transaksi bernilai tinggi. Secara IQR, ambang kandidat outlier berada sekitar >-309k, yang lebih tepat diperlakukan sebagai kasus bisnis bernilai besar daripada kesalahan, namun tetap perlu flagging agar tidak mendominasi proses pembelajaran. Untuk mereduksi heteroskedastisitas dan menormalkan sebaran pada model sensitif, log1p(total_harga) efektif (di ruang log, penjumlahan log-qty dan log-price cenderung lebih mendekati normal).

3.2.4 Korelasi Fitur Numerik



Gambar 5. Korelasi Variabel Numerik

Pada Gambar 5, matriks memperlihatkan korelasi sangat kuat dan positif antara total qty dan total harga (r = 0,98). Secara statistik, pada regresi linier tunggal total harga ~ total qty, nilai ini berimplikasi R^2 = r^2 = 0,96: sekitar 96% variasi total harga dapat dijelaskan oleh perubahan total qty. Hasil ini konsisten dengan relasi deterministik total harga = total qty × harga satuan serta fakta bahwa variasi harga satuan relatif sempit.

Peran harga satuan. Korelasi harga satuan terhadap total harga berada pada level lemah–positif (r = 0,17), sedangkan terhadap total qty hampir nol (r = 0,03). Artinya, unit price relatif independen dari kuantitas pembelian (tidak ada pola diskon kuantitas yang kuat), dan kontribusinya ke variasi nilai transaksi jauh lebih kecil dibandingkan kuantitas.

3.3 Pembersihan Data (*Data Cleaning*)

Pembersihan data dilakukan dengan menghapus nilai kosong (*missing values*) serta duplikasi data menggunakan fungsi dropna() dan drop_duplicates(). Tindakan ini bertujuan untuk menjaga integritas dataset dan mencegah bias model akibat adanya entri yang tidak lengkap atau berulang.

Tabel 5. Ringkasan Pembersihan Data

Fitur	Tipe Interferensi	Misiing_%	Unik
Tgl Order	Datetime64[ns]	0.0	52
Lokasi	Object	0.0	18
Nama Barang	Object	0.0	12

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1649−1661
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8657

Fitur	Tipe Interferensi	Misiing %	Unik
Total Qty	Float 64	0.0	17
Harga Satuan	Float 64	0.0	10
Total Harga	Float 64	0.0	85

Hasil pada Tabel 5, menunjukkan data yang sudah melewati tahapan pembersihan data, tgl order = datetime64[ns], lokasi/nama barang = object, dan metrik (total qty, harga satuan, total harga) = float64; missing% = 0 di semua kolom inti, duplikat = 1 (hapus), dan ketidaksesuaian aritmatika = 0 (aturan bisnis terpenuhi). Artinya dataset bersih dan siap dipakai untuk modelling; lanjutkan dengan label encoding kategori, fitur waktu (bulan/kuartal/dow), dan Avg_Price = Total Harga / Total Qty sebagai fitur tambahan. data dibersihkan dengan;

a. harmonisasi skema (menyatukan nama kolom & tipe: tanggal adalah datetime, numerik adalah float64),
b. sanitasi kategori (lokasi, nama barang) agar tidak memecah kelas saat encoding,
c. validasi domain Total Harga=Total Qty×Harga Satuan dengan toleransi kecil,
d. cek duplikasi dan profiling missing.

3.4 Variabel Target

Penelitian ini diformulasikan sebagai regresi terawasi dengan variabel target kuantitas penjualan (target_qty, turunan terstandar dari “Total qty”). Pemilihan kuantitas bukan nominal memberi sinyal permintaan yang lebih murni dan relevan untuk perencanaan stok, sekaligus mencegah label leakage (karena “Total harga” = “Total qty” × “Harga satuan” maka “Total harga” tidak dipakai sebagai fitur). Fitur prediktor meliputi komponen waktu (bulan/kuartal/hari-dalam-pekan), kategori yang di-encode persisten (lokasi, nama barang), serta harga satuan dan Avg_Price.

3.5 Seleksi Fitur dan Persiapan Data

3.5.1 Seleksi Fitur

Pemilihan fitur dilakukan dengan prinsip anti-leakage, relevansi domain, dan kemudahan generalisasi. Target (label) ditetapkan sebagai kuantitas penjualan (target_qty, turunan terstandar dari “Total qty”). Untuk prediktor (x), dipertahankan fitur yang secara kausal memengaruhi permintaan dan aman dari kebocoran informasi:

a. fitur waktu (bulan, kuartal, hari-dalam-pekan) sebagai penangkap tren dan musiman;
b. fitur kategorikal Lokasi dan Nama barang yang direpresentasikan melalui label-encoding persisten (mapping disimpan agar konsisten saat inferensi);
c. fitur harga Harga satuan serta fitur turunan Avg_Price = Total Harga / Total Qty dengan proteksi pembagi nol untuk mengukur intensitas harga per unit.

Fitur yang berpotensi menyebabkan label leakage terutama “Total Harga” (karena deterministik terhadap target: Total Harga = Total Qty × Harga Satuan) dikecualikan dari set fitur. Korelasi numerik digunakan secara deskriptif (bukan sebagai hard filter) karena model tree-based (XGBoost) robust terhadap multikolinearitas dan tidak memerlukan standarisasi skala.

3.5.2 Persiapan Data

Tahap persiapan data mencakup

a. Harmonisasi skema normalisasi nama kolom (lowercase/trim, pemetaan alias “Harga unit”harga satuan) serta casting tipe: tgl order→datetime64[ns], metrik numerik float64.
b. Sanitasi kategorikal pembersihan spasi ganda/typo untuk mencegah inflasi kardinalitas saat encoding.
c. Validasi domain pengecekan invarian bisnis ,Total Harga=Total Qty×Harga Satuan dengan toleransi numerik kecil.
d. Deduplikasi observasi untuk menjaga estimasi yang tidak bias;
e. Profiling missingness pada data inti tidak ditemukan nilai hilang sehingga imputasi tidak diperlukan;
f. Rekayasa fitur waktu (bulan/kuartal/dow) dan Avg_Price seperti di atas;
g. Pembelahan data menggunakan hold-out 80:20 (atau split temporal bila diarahkan) untuk mencegah temporal leakage.

3.6 Normalisasi Data

Normalisasi pada penelitian ini difokuskan pada standardisasi representasi data agar konsisten, bebas noise format, dan siap dimodelkan bukan pada feature scaling numerik (karena XGBoost tidak memerlukan skala baku). Langkah yang dilakukan mencakup:

a. normalisasi skema: penyelarasan nama kolom (lowercase, trimming, pemetaan alias seperti “Harga unit” harga satuan), serta pengubahan tipe tgl order datetime64[ns] dan metrik (total qty, harga satuan, total harga) float64;
b. normalisasi nilai numerik: pembersihan pemisah ribuan/simbol mata uang sehingga nilai rupiah dan jumlah terbaca sebagai angka murni;
c. normalisasi kategorikal: penghapusan spasi ganda/typo pada lokasi dan nama barang agar kardinalitas kategori tidak terinflasi saat encoding;

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1649−1661
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8657

d. normalisasi semantik/domain: verifikasi invarian bisnis Total Harga=Total Qty×Harga Satuan dengan toleransi numerik kecil, sehingga hanya nilai yang konsisten yang diteruskan ke pemodelan. Dengan desain ini, skala asli variabel dipertahankan untuk menjaga interpretabilitas metrik

3.6 Encoder

Penelitian ini menerapkan skema hold-out dengan proporsi 80:20 untuk membentuk himpunan pelatihan (df_train) dan pengujian (df_test) dari dataset yang telah dinormalisasi. Setelah harmonisasi skema (penyeragaman nama kolom dan tipe data) dan rekayasa fitur waktu (bulan, kuartal, hari-dalam-pekan) serta fitur intensitas harga (Avg_Price), dilakukan seleksi fitur anti-leakage, variabel target ditetapkan sebagai kuantitas penjualan (turunan dari "Total qty"), sementara "Total harga" tidak digunakan sebagai prediktor karena deterministik terhadap target (Total harga = Total qty × Harga satuan). Dua atribut kategorikal utama Lokasi dan Nama barang kemudian direpresentasikan secara numerik melalui Ordinal/Label Encoding persisten (handle_unknown = -1) yang difit hanya pada data latih dan disimpan sebagai berkas model encoder. Pendekatan ini memastikan konsistensi representasi lintas siklus (train,test,deploy) serta mencegah category drift ketika dilakukan inferensi pada periode berikutnya, yaitu tahun 2026.

3.7 Implementasi Model XGBoost

```
model = XGBRegressor(
    n_estimators=300,
    learning_rate=0.1,
    max_depth=6,
    random_state=42,
    subsample=0.8,
    colsample_bytree=0.8
)
model.fit(X_train, y_train)
```

```
XGBRegressor
XGBRegressor(base_score=None, booster=None, callbacks=None,
             colsample_bylevel=None, colsample_bynode=None,
             colsample_bytree=0.8, device=None, early_stopping_rounds=None,
             enable_categorical=False, eval_metric=None, feature_types=None,
             feature_weights=None, gamma=None, grow_policy=None,
             importance_type=None, interaction_constraints=None,
             learning_rate=0.1, max_bin=None, max_cat_threshold=None,
             max_cat_to_onehot=None, max_delta_step=None, max_depth=6,
             max_leaves=None, min_child_weight=None, missing=nan,
             monotone_constraints=None, multi_strategy=None, n_estimators=300,
             n_jobs=None, num_parallel_tree=None, ...)
```

Gambar 7. Implementasi Model XGBoost

Gambar 7, Implementasi Model XGBoost. Setelah dilakukan Data Preparaton/Persiapan Data selanjutnya adalah tahapan pelatihan model. Dalam penelitian ini menggunakan model XGBoost, model ini dipilih karena tingkat akurasi nya yang tinggi dalam memprediksi sebuah data. Model XGBoost dilatih menggunakan data hasil data preparation dengan pembagian data latih sebesar 80 persen dan data uji sebesar 20 persen. Parameter yang digunakan dalam proses pelatihan meliputi learning rate = 0.1, max_depth = 6, dan n_estimators = 100, dengan fungsi tujuan (objective function) berupa reg:squarederror. Pemilihan parameter tersebut dilakukan untuk mencapai keseimbangan antara kecepatan pelatihan dan akurasi hasil prediksi. Secara umum, proses pelatihan dilakukan melalui pemanggilan fungsi fit() pada pustaka XGBoost, di mana model membangun sejumlah pohon keputusan secara bertahap. Setiap pohon yang baru dibangun berupaya memperbaiki kesalahan yang dihasilkan oleh pohon sebelumnya, sehingga model akhir merupakan kombinasi beberapa pohon yang memberikan hasil prediksi paling optimal. Setelah proses pelatihan selesai, model digunakan untuk melakukan prediksi terhadap data uji menggunakan fungsi predict(). Hasil prediksi kemudian dibandingkan dengan nilai aktual guna menilai tingkat akurasi model. Evaluasi kinerja dilakukan menggunakan dua metrik utama, yaitu Mean Absolute Error (MAE) dan Koefisien Determinasi (R^2).

3.8 Evaluasi Model

Pada tahap evaluasi, kinerja model XGBoost dinilai menggunakan dua metrik utama, yaitu Mean Absolute Error (MAE) dan koefisien determinasi (R^2) sebagaimana ditunjukkan pada Gambar 8. Nilai MAE yang dihasilkan sebesar 0,0549, sedangkan nilai R^2 mencapai 0,9936. Kombinasi kedua metrik ini memberikan gambaran bahwa model mampu melakukan prediksi penjualan kue kering dengan tingkat kesalahan yang relatif sangat kecil dan kemampuan penjelasan variasi data yang sangat tinggi.

```
=== EVALUASI MODEL XGBOOST ===
MAE   : 0.0549
R²    : 0.9936
```

Gambar 8. Evaluasi Model

3.8.1 MAE (Mean Absolute Error)

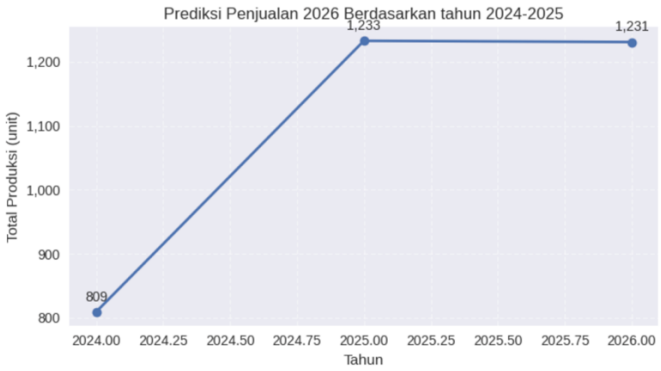
Pada gambar 8, hasil evaluasi menghasilkan MAE = 0,0549, yang menunjukkan besaran rata-rata deviasi absolut antara nilai aktual dan prediksi pada skala target yang digunakan. Jika target tidak mengalami transformasi, nilai ini setara dengan sekitar 0,055 unit per observasi, sehingga secara operasional mengindikasikan tingkat kesalahan yang sangat kecil pada prakiraan kuantitas penjualan. Capaian ini konsisten dengan pipeline anti-leakage dan kualitas data yang bersih (tanpa missing pada kolom inti, konsistensi aritmatika), serta pemilihan fitur yang relevan (waktu, kategori ter-encode persisten, harga satuan, dan Avg_Price).

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1649−1661
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8657

3.8.2 R^2(Koefisiensi Determinasi)

Pada Gambar 8, nilai R^2 = 0,9936 menunjukkan bahwa model menjelaskan sekitar 99,36% variansi target pada data uji, menandakan kesesuaian yang sangat tinggi antara prediksi dan data observasi. Secara metodologis, hal ini menegaskan bahwa struktur pola yang ditangkap model terutama musiman/temporal dan heterogenitas lokasi–produk telah terwakili dengan baik oleh fitur yang dibangun. Dengan demikian, model memiliki daya jelaskan yang kuat dan berpeluang tinggi untuk mendukung keputusan operasional (mis. perencanaan stok dan replenishment), dengan tetap dianjurkan pemantauan residual dan evaluasi per segmen untuk memastikan kinerja yang adil di seluruh cabang dan varian produk.

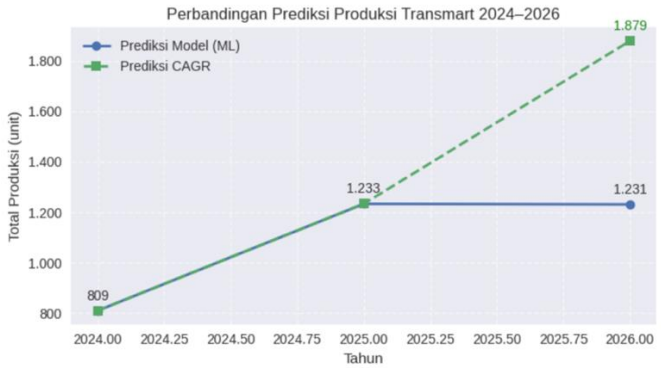
3.9 Prediksi Model



Gambar 9. Prediksi Model

Grafik Pada Gambar 9 menampilkan tren total produksi/penjualan periode 2024–2026, dengan nilai aktual pada 2024 sebesar 809 unit dan peningkatan tajam pada 2025 menjadi 1.233 unit. Titik 2026 merupakan hasil prediksi model XGBoost yang menunjukkan stabilisasi pada kisaran 1.231 unit. Kenaikan kuat 2024-2025 diikuti plateau 2025-2026 (selisih 2 unit) mengindikasikan pasar memasuki fase pemantapan setelah ekspansi awal. Secara operasional, pola ini menyarankan penyesuaian perencanaan stok menuju level yang lebih konservatif pada 2026, pengendalian produksi berlebih, serta fokus pada cabang/varian berkontribusi tinggi agar utilisasi persediaan tetap efisien. Dari sisi pemodelan, keluaran XGBoost menangkap dinamika non-linier tanpa asumsi pertumbuhan linier; oleh karena itu, proyeksi ini relevan sebagai dasar uji skenario (mis. perubahan harga/promosi) dan penetapan target penjualan tahunan.

3.10 Perbandingan Performa Model dengan CAGR



Gambar 10. Perbandingan Model vs CAGR

Gambar 10 menunjukkan perbandingan hasil prediksi total produksi Transmart untuk periode 2024 hingga 2026 yang diperoleh dari tiga metode berbeda, yaitu model Machine Learning (ML) berbasis XGBoost, metode matematis Compound Annual Growth Rate (CAGR), dan model Hybrid yang merupakan gabungan dari kedua metode tersebut. Grafik dalam gambar ini disajikan dalam bentuk grafik garis (line chart), di mana sumbu horizontal (X) menunjukkan tahun observasi (2024, 2025, dan 2026), sedangkan sumbu vertikal (Y) menunjukkan jumlah total produksi atau penjualan dalam satuan unit.

Dari gambar tersebut, terlihat bahwa model ML memproyeksikan total produksi meningkat tajam, dari 809 unit pada tahun 2024 menjadi 1.233 unit pada tahun 2025, namun sedikit menurun ke 1.230 unit pada tahun 2026. Pola ini menunjukkan bahwa produksi mulai stabil setelah mencapai puncak pertumbuhan pada tahun sebelumnya. Sementara itu, metode CAGR menunjukkan peningkatan yang lebih tinggi, dengan tren terus meningkat hingga mencapai 1.879 unit pada tahun 2026, menandakan proyeksi yang lebih optimis dengan pertumbuhan tahunan sebesar

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1649−1661
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8657

52,41%. Model Hybrid, yang merupakan kombinasi dari kedua metode tersebut, menunjukkan hasil yang lebih moderat, dengan prediksi total produksi sebesar 1.555 unit pada tahun 2026 dan tingkat pertumbuhan sebesar 26,12%, secara visual berada di antara dua garis prediksi lainnya. Secara analitis, hasil ini menunjukkan perbedaan karakteristik antara pendekatan berbasis data (ML) dan pendekatan matematis (CAGR).

4. KESIMPULAN

Penelitian ini menunjukkan bahwa menggunakan algoritme XGBoost dalam kerangka CRISP-DM sangat efektif untuk membuat model dan memprediksi permintaan belanja ritel berdasarkan data transaksi tahun 2024 hingga 2025. Setelah melakukan analisis awal data (EDA), membersihkan data (tanpa ada nilai kosong pada kolom utama), merekayasa fitur berbasis waktu, serta melakukan label encoding pada variabel kategorikal (lokasi dan nama produk), model dibangun dengan skema pembagian data 80:20. Dari hasil evaluasi pada data uji, diperoleh MAE = 0,0549 dan R^2 = 0,9936, yang menunjukkan akurasi sangat tinggi serta kemampuan model dalam menggambarkan perubahan target secara baik. Prediksi secara keseluruhan menunjukkan peningkatan dari 809 unit pada tahun 2024 menjadi 1.233 unit pada 2025, lalu stabil hingga sekitar 1.230 unit pada 2026 dengan penurunan kecil sekitar 0,17%. Temuan ini mengindikasikan bahwa pasar mulai memasuki fase stabil, sehingga strategi perencanaan stok yang lebih hati-hati dan disesuaikan dengan segmen lokasi serta produk menjadi lebih relevan. Dibandingkan dengan pendekatan matematis yang menggunakan CAGR (tahunan pertumbuhan konstan), pendekatan XGBoost lebih realistis dalam memperhitungkan dinamika yang tidak linier dan perbedaan antar lokasi serta produk. Sementara itu, model hibrida menjadi pilihan yang seimbang ketika diperlukan proyeksi yang tidak terlalu optimistis tetapi tetap bisa menanggapi perubahan. Secara manajerial, model ini sangat berguna untuk menyiapkan persediaan, mendistribusikan barang antar toko, serta mengevaluasi kebijakan harga atau promosi berdasarkan simulasi.

REFERENCES

[1] W. Wardiana, “Perkembangan Teknologi Informasi di Indonesia,” *Seminar dan Pameran Teknologi Informasi*, 2002.
[2] K. Hadiono dan R. C. N. Santi, “Menyongsong Transformasi Digital,” *Proceeding SENDIU* , hlm. 81–84, 2020.
[3] B. W. Sari dan D. Prabowo, “Analisis Perbandingan Prediksi Harga Rumah Dengan Random Forest, Gradient Boosting, dan XGBoost,” *Intellect : Indonesian Journal of Innovation Learning and Technology*, vol. 4, no. 1, hlm. 42–51, 2025, doi: 10.57255/intellect.v4i1.1385.
[4] M. A. Rayadin, M. Musaruddin, R. A. Saputra, dan Isnawaty, “Implementasi Ensemble Learning Metode XGBoost dan Random Forest untuk Prediksi Waktu Penggantian Baterai Aki,” *BIOS: Jurnal Teknologi Informasi dan Rekayasa Komputer*, vol. 5, no. 2, hlm. 111–119, 2024, doi: 10.37148/bios.v5i2.128.
[5] A. A. Saputra, B. N. Sari, dan C. Rozikin, “Penerapan Algoritma Extreme Gradient Boosting (Xgboost) Untuk Analisis Risiko Kredit,” *Jurnal Ilmiah Wahana Pendidikan*, vol. 10, no. 7, hlm. 27–36, 2024, doi: 10.5281/zenodo.10960080.
[6] F. L. Febrianti, I. M. Nur, A. M. Haris, dan S. Amri, “Implementasi Metode XGBoost dan SHAP untuk Klasifikasi dan Analisis Faktor Risiko Penyakit Diabetes Mellitus,” *Seminar Nasional Sains Data (SENADA)*, vol. 2025, hlm. 336–346, 2025.
[7] P. M. Izzati dan Fitriyani, “Implementasi Algoritma XGBoost Untuk Prediksi Capaian Bulanan Pendapatan Daerah Kota Bandung,” *Jurnal Computer Science and Information Technology (CoSciTech)*, vol. 6, no. 2, hlm. 104–111, 2025, doi: 10.37859/coscitech.v6i2.9578.
[8] B. Mardika, S. Utami, dan J. Widiyanto, “Identifikasi Keanekaragaman Gastropoda Kualitas Air Sungai Nogosari Pacitan,” *Prosiding Seminar Nasional Simbiosis*, hlm. 349–357, 2020.
[9] A. N. Hidayat, “Implementasi XGBoost dalam Klasifikasi Gagal Ginjal Kronis Menggunakan Dataset Chronic Kidney Disease,” *Jurnal Teknik Informatika dan Sistem Informasi*, 2025.
[10] S. H. Sinurat, “Analisis Big Data Dengan Metode Exploratory Data Analysis (Eda) Dan Metode Visualisasi Menggunakan Jupyter Notebook,” *Jurnal Sistem Informasi dan Ilmu Komputer Prima (JUSIKOM PRIMA)*, vol. 4, no. 2, 2021.
[11] M. A. Hasanah, S. Soim, dan A. S. Handayani, “Implementasi CRISP-DM Model Menggunakan Metode Decision Tree dengan Algoritma CART untuk Prediksi Curah Hujan Berpotensi Banjir,” *Journal of Applied Informatics and Computing (JAIC)*, vol. 5, no. 2, hlm. 103–108, 2021.
[12] J. Syam, “Compound Annual Growth Rate (CAGR) untuk Menganalisis Tren Populasi Sapi Potong di Kabupaten Pinrang,” *Jurnal Sains dan Teknologi Industri Peternakan*, vol. 5, no. 2, 2025, doi: 10.55678/jstip.v5i2.2210.
[13] F. Latuni, S. Gampu, dan Y. P. Yusup, “Analisis Tren Laporan Laba Rugi PT Astra International Tbk Tahun (2013-2023),” *Global Science*, vol. 5, no. 1, 2024.
[14] R. Haris, “Analisis Tren Pasar dan Pengambilan Keputusan Berbasis Data dalam Meningkatkan Daya Saing Bisnis,” *ADI Bisnis Digital Interdisiplin (ABDI Jurnal)*, 2024.
[15] R. Winurputra dan D. E. Ratnawati, “Peramalan Penjualan Produk Menggunakan Extreme Gradient Boosting (XGBoost) dan Kerangka Kerja CRISP-DM untuk Pengoptimalan Manajemen Persediaan (Studi Kasus: UB Mart),” *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)*, vol. 12, no. 2, hlm. 417–428, 2025, doi: 10.25126/jtiik.2025129451.
[16] M. Abdillah *dkk.*, “Implementasi XGBoost dalam Klasifikasi Gagal Ginjal Kronis Menggunakan Dataset Chronic Kidney Disease,” *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 12, no. 3, 2025, [Daring]. Tersedia pada: http://jurnal.mdp.ac.id
[17] N. N. P. Pinata, I. M. Sukarsa, dan N. K. D. Rusjayanthi, “Prediksi Kecelakaan Lalu Lintas di Bali dengan XGBoost pada Python,” *Jurnal Ilmiah Merpati*, vol. 8, no. 3, 2020.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1649−1661
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8657

[18] Z. B. P. Pratama dan Y. P. Astuti, "Perbandingan Metode Machine Learning (Linear Regression, Random Forest, dan XGBoost) dalam Memprediksi Kemiskinan di Jawa Tengah Tahun 2024," *Sistemasi: Jurnal Sistem Informasi*, vol. 14, no. 5, 2025, [Daring]. Tersedia pada: http://sistemasi.ftik.unisi.ac.id

[19] A. Cahyana, E. R. Susanto, dan Parjito, "Penerapan Algoritma XGBoost untuk Prediksi Diabetes: Analisis Confusion Matrix dan ROC Curve," *Fountain of Informatics Journal*, vol. 10, no. 1, hlm. 40–50, 2025, doi: 10.21111/fij.v10i1.14311.

[20] H. Wijaya, P. D. Hostiadi, dan E. Triandini, "Meningkatkan Prediksi Penjualan Retail Xyz dengan Teknik Optimasi Random Search pada Model Xgboost," *Prosiding Seminar Hasil Penelitian Informatika dan Komputer (SPINTER)*, vol. 1, no. 2, hlm. 829–833, 2024.

[21] A. N. Rachmi, "Implementasi Metode Random Forest dan XGBoost pada Klasifikasi Customer Churn," *Universitas Islam Indonesia Yogyakarta*, 2020.

[22] L. W. Maahiroh, "Klasifikasi Turnover Karyawan Menggunakan Algoritma XGBoost (Studi kasus: Divisi Engineering, Perusahaan Jasa Pertambangan)," *Universitas Islam Indonesia Yogyakarta*, 2024.

[23] S. F. Sihombing, J. P. Pakpahan, dan A. H. Lubis, "Klasifikasi Produk Iphone dengan Menggunakan Algoritma XGBoost," *Journal of Informatics Management and Information Technology*, vol. 5, no. 3, hlm. 371–380, 2025, doi: 10.47065/jimat.v5i3.649.

[24] M. F. Reza, "Implementasi Model Gated Recurrent Unit (GRU) atau Extreme Gradient Boosting (XGBoost) untuk Prediksi Harga Cryptocurrency Ethereum," *Jurnal Rekayasa Sistem Informasi dan Teknologi (JRSIT)*, vol. 3, no. 1, 2025,

[25] M. K. K. Ichwanul, "Implementasi Metode XGBoost dan Feature Importance untuk Klasifikasi pada Kebakaran Hutan dan Lahan," *Journal of Software Engineering, Information and Communication Technology*, vol. 1, no. 1, hlm. 11–18, 2020.

[26] M. Salsabil, N. L. Azizah, dan A. Eviyanti, "Implementasi Data Mining dalam Melakukan Prediksi Penyakit Diabetes Menggunakan Metode Random Forest dan Xgboost," *Jurnal Ilmiah Komputasi*, vol. 23, no. 1, hlm. 51–58, 2024, doi: 10.32409/jikstik.23.1.3507.

[27] S. A. Tiastama dan I. Budi, "Perbandingan Random Search dan Algoritma Genetika dalam Penyetelan Hyperparameter XGBoost pada Retail Sales Forecasting," *The Indonesian Journal of Computer Science*, vol. 13, no. 4, hlm. 6602–6613, 2024, doi: 10.33022/ijcs.v13i4.4285.

[28] R. Abdurrosyid dan A. T. W. Almais, "Deteksi Dini Diabetes menggunakan Machine Learning dengan Metode PCA dan XGBoost," *Jurnal Edukasi dan Penelitian Informatika*, vol. 11, no. 1, hlm. 51–56, 2025.

[29] I. G. A. R. Astarani dan I. G. S. Rahayuda, "Analisis Perbandingan XGBoost dan LightGBM dalam Prediksi Penjualan Ritel Walmart Store Sales," *Jurnal Nasional Teknologi Informasi dan Aplikasinya (JNATIA)*, vol. 3, no. 4, hlm. 717–728, 2025.

[30] M. M. Ibrahim, "Analisis Kinerja Model Machine Learning untuk Mendeteksi Transaksi Fraud pada Sistem Pembayaran Online," *Jurnal Ilmiah Nusantara (JINU)*, vol. 2, no. 3, hlm. 35–49, 2025, doi: 10.61722/jinu.v2i3.4276.

[31] A. P. F. Prasetya dan P. H. P. Rosa, "Klasifikasi Kegagalan Pembayaran Kredit Nasabah Bank dengan Algoritma XGBoost," *Seminar Nasional Informatika Bela Negara (SANTIKA)*, no. 4, hlm. 366–371, 2024, [Daring]. Tersedia pada: https://www.kaggle.com/datasets/nikhil1e9/loan-default/data.