Journal of Innovative and Creativity
Journal Homepage: https://joecy.org/index.php/joecy
e-mail: galungwaknook@gmail.co

Faculty of Education
University of Pahlawan Tuanku Tambusai
ISSN 2775-771X (Online)
ISSN 2962-570X (Printed)
6 (2) 2026, Page: 27798-27812

Penerapan Data Mining Dalam Prediksi Penjualan Menggunakan Algoritma Klasifikasi dan Regresi : Tinjauan Pustaka

Rahmad Winardi[1], Nurdin[2]

[1,2]Universitas Malikussaleh
Corresponding Author: e-mail: rahmad.257110201028@mhs.unimal.ac.id[1], nurdin@unimal.ac.id[2]

Published:, Mei 2026

ABSTRACT

This study aims to analyze the application of data mining in sales forecasting using classification and regression algorithms through a systematic literature review and case study approach. The review examined 25 included articles published in 2020 or later, selected through a PRISMA-based screening process. The analysis focused on the types of algorithms used, the format of sales data, evaluation metrics, key results, and research gaps. The results indicate that regression algorithms are the most widely used approach because sales predictions are generally aimed at estimating sales value, demand volume, revenue, or product volume. Random Forest Regression, XGBoost, Linear Regression, Support Vector Regression, Gradient Boosting, LSTM, ARIMA, and Neural Networks are frequently employed in this category. Classification algorithms are used to determine sales categories, such as best-selling and slow-moving products, purchase likelihood, and sales performance categories. Random Forest and XGBoost are algorithms that frequently demonstrate strong results on retail and e-commerce data. Hybrid models are also increasingly being applied through the combination of RFM, Apriori, sentiment analysis, and regression models. Research findings indicate that the success of sales predictions is determined not only by the algorithms but also by data quality, feature selection, model validation, and the model's alignment with business objectives.

Keywords: *data mining; sales forecasting; classification algorithms; regression algorithms*

© The Author(s). 2021 Open Access This article is licensed under a Creative Commons Atibution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0

PENDAHULUAN

Perubahan pola belanja digital telah membuat data penjualan menjadi sumber informasi yang sangat penting bagi pelaku usaha. Setiap transaksi menyimpan jejak perilaku konsumen. Data produk, harga, waktu pembelian, lokasi, promosi, kanal penjualan, dan metode pembayaran dapat dibaca sebagai pola permintaan. Urgensi ini makin kuat setelah nilai transaksi e-commerce Indonesia tahun 2024 mencapai Rp1.288,93 triliun dan meningkat 17,08% dibandingkan tahun 2023. Transaksi tersebut paling besar berasal dari media non-marketplace sebesar 84,21%. Bank Indonesia juga mencatat volume transaksi pembayaran digital pada triwulan III 2025 mencapai 12,99 miliar transaksi atau tumbuh 38,08% secara tahunan. Angka tersebut memperlihatkan bahwa aktivitas belanja digital menghasilkan data besar yang tidak cukup hanya dibaca melalui laporan penjualan biasa. Pelaku usaha memerlukan model prediksi yang lebih terukur agar keputusan stok, promosi, penetapan harga, dan distribusi dapat dilakukan lebih tepat (Kementerian Perdagangan, 2026; Bank Indonesia, 2025).

Prediksi penjualan menjadi bagian penting dari data mining karena data historis dapat diolah menjadi informasi yang membantu pengambilan keputusan bisnis. Algoritma klasifikasi dapat dipakai untuk menentukan kelas penjualan, seperti tinggi, sedang, rendah, laku, tidak laku, atau berisiko habis stok. Algoritma regresi dapat digunakan untuk memperkirakan nilai penjualan pada periode tertentu. Dua pendekatan ini memiliki peran yang berbeda. Klasifikasi lebih dekat dengan keputusan kategoris. Regresi lebih dekat dengan proyeksi angka. Prabu et al. (2025) membahas ragam model prediksi penjualan seperti LightGBM, XGBoost, Random Forest, Decision Tree, KNN, Ridge Regression, Elastic Net, dan Linear Regression dengan metrik evaluasi MAE, RMSE, serta R-squared. Eglite dan Birzniece (2022) menempatkan akurasi prediksi penjualan ritel sebagai dasar penting untuk menjaga stok, biaya penyimpanan, dan kelancaran promosi. Makridakis et al. (2022) juga memperlihatkan

bahwa kompetisi M5 menggunakan 42.840 deret waktu penjualan ritel Walmart, termasuk variabel penjelas, sehingga kebutuhan model prediksi tidak lagi terbatas pada pola waktu sederhana.

Riset terdahulu memperlihatkan bahwa algoritma regresi dan ensemble learning sering dipilih karena mampu menangkap hubungan antara penjualan dan variabel pendukung. Massaro et al. (2021) memakai augmented data dan XGBoost untuk meningkatkan prediksi penjualan. Panarese et al. (2022) mengembangkan platform prediksi penjualan berbasis gradient boosting dan menekankan pentingnya fitur yang relevan bagi kebutuhan bisnis. Haselbeck et al. (2022) membandingkan model klasik dan machine learning pada penjualan hortikultura. XGBoost menjadi model terbaik pada 14 dari 15 perbandingan. Ensafi et al. (2022) membandingkan model machine learning untuk penjualan barang musiman. Nasseri et al. (2023) membandingkan tree-based ensemble dengan deep learning berbasis LSTM untuk prediksi permintaan ritel. Temuan-temuan tersebut memberi arah bahwa prediksi penjualan tidak hanya bergantung pada model statistik dasar. Akurasi sering meningkat saat data promosi, musim, kalender, harga, serta pola permintaan jangka pendek ikut dipakai sebagai fitur prediktor.

Pendekatan klasifikasi juga memiliki posisi penting karena tidak semua keputusan penjualan membutuhkan angka pasti. Banyak keputusan bisnis justru dimulai dari pengelompokan. Produk dapat diklasifikasikan sebagai laris atau tidak laris. Pelanggan dapat diklasifikasikan menurut potensi pembelian. Transaksi dapat diklasifikasikan menurut peluang keberhasilan penjualan. Purwanto et al. (2026) membandingkan Decision Tree, Random Forest, SVM, Naive Bayes, dan KNN untuk prediksi penjualan e-commerce. Random Forest mencapai akurasi 97,5% dan AUC 0,991. Temuan ini memperlihatkan bahwa klasifikasi dapat membantu pemetaan hasil penjualan secara cepat. Model klasifikasi juga lebih mudah dipakai saat manajemen membutuhkan label keputusan. Nilai prediksi yang berbentuk kelas sering lebih mudah diterjemahkan menjadi strategi, seperti prioritas promosi, segmentasi produk, atau peringatan stok.

Persoalan lain muncul saat produk baru belum memiliki data historis yang panjang. Hwang et al. (2023) mengembangkan model prediksi penjualan untuk produk baru dan produk berdaur hidup pendek pada kasus telepon seluler. Elalem et al. (2023) mengusulkan kerangka machine learning untuk meramalkan penjualan produk baru yang memiliki kemiripan dengan produk terdahulu. Hwang et al. (2025) memperkuat arah tersebut melalui penggabungan homogeneity-based clustering dan ensemble learning agar prediksi produk baru menjadi lebih akurat. Tang dan Dong (2021) juga membahas prediksi total penjualan produk baru berdaur hidup pendek melalui penggabungan informasi multi-sumber. Rangkaian riset ini memberi gambaran bahwa prediksi penjualan memerlukan pemilihan algoritma yang sesuai dengan karakter data. Data panjang dapat diarahkan ke regresi deret waktu atau ensemble. Data terbatas dapat dibantu oleh pengelompokan, kemiripan produk, dan klasifikasi berbasis fitur.

Kajian berbasis systematic literature review diperlukan karena penelitian prediksi penjualan masih tersebar pada berbagai sektor, algoritma, metrik, dan sumber data. Douaioui et al. (2024) menelaah 119 artikel Scopus tahun 2015–2024 mengenai machine learning dan deep learning untuk demand forecasting pada manajemen rantai pasok. Saptadi et al. (2026) menyeleksi 66 publikasi dari 206 publikasi awal mengenai machine learning untuk sales prediction. XGBoost dan Random Forest disebut sebagai algoritma yang sering digunakan dan konsisten lebih unggul dibandingkan regresi tradisional. Dua tinjauan tersebut memberi dasar kuat bahwa area prediksi penjualan berkembang cepat. Akan tetapi, peta khusus yang membandingkan peran algoritma klasifikasi dan regresi pada prediksi penjualan masih perlu dipertegas. Banyak artikel menonjolkan performa model. Pembahasan mengenai jenis keputusan bisnis yang cocok untuk klasifikasi dan jenis keputusan yang cocok untuk regresi belum selalu disajikan secara runtut.

Pemilihan algoritma juga perlu dibaca dari pengalaman penerapan data mining pada bidang lain. Nurdin et al. (2022) membandingkan Naive Bayes dan Dempster Shafer untuk diagnosis awal COVID-19 dan memperoleh akurasi Naive Bayes sebesar 96%. Salamah et al. (2024) membandingkan K-Means dan K-Medoids untuk penentuan program studi dengan evaluasi Davies-Bouldin Index. Salsabila et al. (2025) memakai K-Means dan K-Medoids untuk pengelompokan wilayah perikanan tangkap di Aceh. Tiga artikel tersebut memberi penegasan bahwa performa algoritma sangat bergantung pada bentuk data, tujuan analisis, serta ukuran evaluasi yang dipakai. Prinsip ini relevan

bagi prediksi penjualan. Algoritma yang baik pada satu jenis data belum tentu unggul pada data lain. Penelitian prediksi penjualan perlu menempatkan klasifikasi, regresi, clustering pendukung, pemilihan fitur, serta evaluasi akurasi sebagai satu alur analisis yang jelas (Nurdin et al., 2022; Salamah et al., 2024; Salsabila et al., 2025).

Celah penelitian tampak pada belum adanya sintesis yang secara khusus membaca penerapan data mining untuk prediksi penjualan melalui dua jalur utama, yaitu klasifikasi dan regresi. Literatur terdahulu banyak menilai algoritma dari sisi akurasi. Pembahasan tentang kesesuaian algoritma dengan tujuan bisnis masih perlu diperluas. Klasifikasi perlu ditelaah untuk kebutuhan penentuan kelas penjualan, segmentasi produk, dan status permintaan. Regresi perlu ditelaah untuk kebutuhan estimasi nilai penjualan, jumlah unit, serta proyeksi pendapatan. SLR pada topik ini dapat menghubungkan jenis data, teknik praproses, algoritma, metrik evaluasi, dan hasil prediksi secara lebih tertata. Penelitian ini juga penting bagi pelaku usaha yang mulai memiliki data transaksi besar, tetapi belum memiliki pedoman yang jelas untuk memilih model prediksi penjualan. Tujuan penelitian ini adalah menganalisis secara sistematis penerapan data mining pada prediksi penjualan melalui algoritma klasifikasi dan regresi

METODE

Penelitian ini menggunakan desain systematic literature review (SLR) yang dipadukan dengan studi kasus untuk menganalisis penerapan data mining pada prediksi penjualan menggunakan algoritma klasifikasi dan regresi. SLR dipilih karena penelitian ini bertujuan menelusuri, menyeleksi, menilai, dan menyintesis artikel ilmiah secara sistematis agar diperoleh peta pengetahuan mengenai algoritma yang digunakan, jenis data penjualan, teknik praproses data, metrik evaluasi, hasil prediksi, serta celah penelitian yang masih dapat dikembangkan. Studi kasus digunakan sebagai pelengkap untuk memberi gambaran penerapan algoritma klasifikasi dan regresi pada data penjualan secara lebih nyata. Pelaporan SLR mengacu pada pedoman PRISMA 2020 yang menjelaskan alur identifikasi, penyaringan, penilaian kelayakan, dan penetapan artikel inklusi (Page et al., 2021).

Penelitian ini difokuskan pada artikel ilmiah tahun 2020 ke atas yang membahas prediksi penjualan, peramalan penjualan, data mining, machine learning, algoritma klasifikasi, dan algoritma regresi. Artikel yang dikaji tidak digabungkan secara statistik karena penelitian ini bukan penelitian meta-analisis. Analisis dilakukan secara deskriptif dan tematik dengan membaca pola temuan dari setiap artikel. Fokus utama diarahkan pada bagaimana algoritma klasifikasi digunakan untuk memprediksi kelas penjualan seperti tinggi, sedang, rendah, laris, tidak laris, atau berisiko menurun. Fokus lain diarahkan pada bagaimana algoritma regresi digunakan untuk memprediksi nilai penjualan, jumlah permintaan, omzet, atau volume produk pada periode tertentu.

Tahapan penelitian meliputi: (1) perumusan fokus SLR, (2) penyusunan kata kunci dan strategi pencarian, (3) pencarian artikel pada basis data ilmiah, (4) seleksi artikel sesuai kriteria inklusi dan eksklusi, (5) penilaian kelayakan full-text, (6) ekstraksi data dari artikel inklusi, (7) sintesis deskriptif dan tematik, serta (8) penyusunan studi kasus penerapan algoritma klasifikasi dan regresi pada data penjualan. Pencarian artikel dilakukan pada empat basis data utama, yaitu Scopus, IEEE Xplore, ScienceDirect, dan Google Scholar. Artikel yang digunakan merupakan artikel peer-reviewed, berbahasa Indonesia atau Inggris, terbit tahun 2020 ke atas, serta membahas penerapan data mining atau machine learning untuk prediksi penjualan.

Tabel 1. Kriteria Inklusi dan Eksklusi Penelitian

Kriteria	Inklusi	Eksklusi
Tahun publikasi	Artikel terbit tahun 2020 ke atas	Artikel terbit sebelum tahun 2020
Topik	Data mining, machine learning, prediksi penjualan, sales prediction, sales forecasting	Artikel yang tidak membahas prediksi penjualan
Algoritma	Memuat algoritma klasifikasi dan/atau regresi	Artikel yang hanya membahas sistem informasi tanpa model prediksi

Jenis publikasi	Artikel jurnal dan prosiding ilmiah peer-reviewed	Skripsi, tesis, editorial, opini, blog, dan artikel populer
Bahasa	Bahasa Indonesia atau Inggris	Bahasa selain Indonesia dan Inggris
Akses teks	Full-text tersedia	Full-text tidak tersedia
Jenis analisis	Memuat proses data, algoritma, evaluasi, atau hasil prediksi	Tidak menjelaskan metode atau hasil prediksi secara jelas

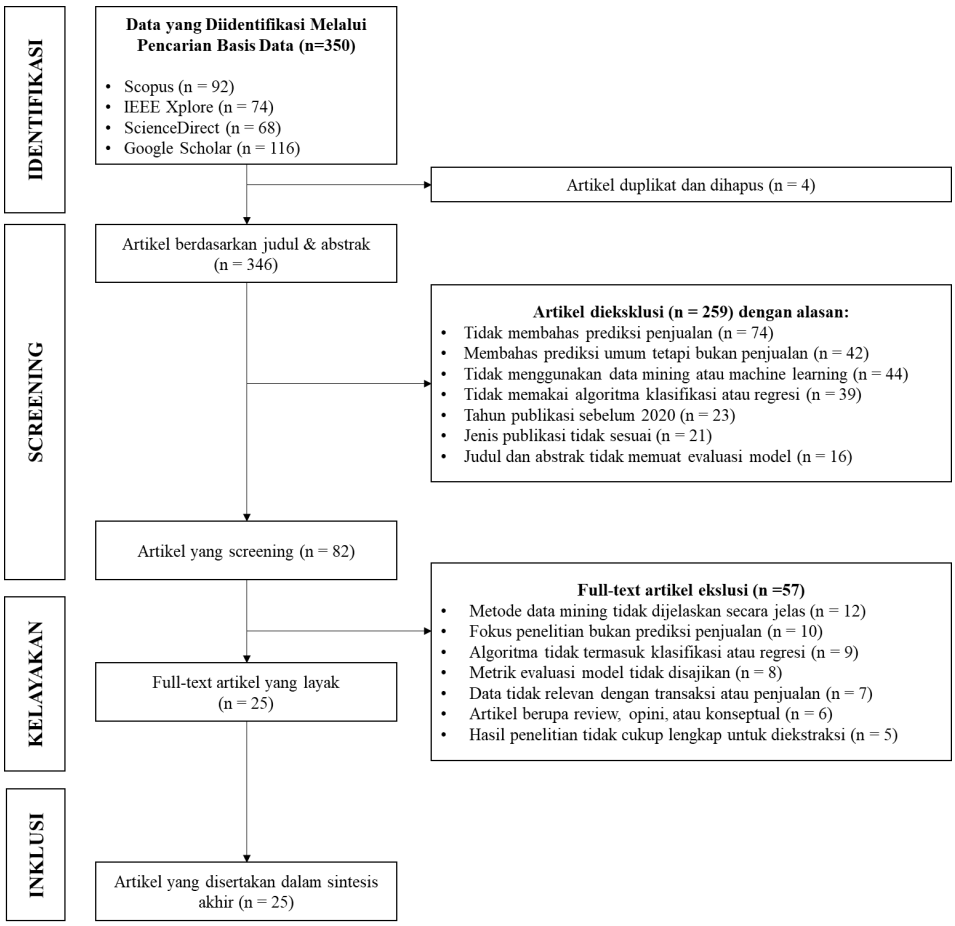
Kata kunci dikembangkan dari istilah utama penelitian, yaitu data mining, prediksi penjualan, klasifikasi, regresi, machine learning, sales prediction, dan sales forecasting. Operator Boolean AND digunakan untuk menghubungkan konsep utama. Operator OR digunakan untuk memperluas variasi istilah yang memiliki makna berdekatan. Strategi pencarian disesuaikan dengan karakter setiap basis data.

Tabel 2. Jumlah Artikel berdasarkan Basis Data Pencarian

Basis Data	Keyword	Jumlah Artikel
Scopus	TITLE-ABS-KEY(("data mining" OR "machine learning") AND ("sales prediction" OR "sales forecasting") AND ("classification" OR "regression"))	92
IEEE Xplore	("data mining" OR "machine learning") AND ("sales prediction" OR "sales forecasting") AND ("classification algorithm" OR "regression algorithm")	74
ScienceDirect	("sales prediction" OR "sales forecasting") AND ("data mining" OR "machine learning") AND ("classification" OR "regression")	68
Google Scholar	"data mining" "prediksi penjualan" "klasifikasi" "regresi" OR "sales prediction" "classification" "regression"	116
Total artikel teridentifikasi		**350**

Proses pencarian artikel pada Scopus, IEEE Xplore, ScienceDirect, dan Google Scholar menghasilkan 350 artikel. Artikel tersebut terdiri atas 92 artikel dari Scopus, 74 artikel dari IEEE Xplore, 68 artikel dari ScienceDirect, dan 116 artikel dari Google Scholar. Setelah dilakukan pemeriksaan duplikasi, ditemukan 4 artikel duplikat. Artikel duplikat tersebut dikeluarkan, sehingga tersisa 346 artikel untuk tahap screening judul dan abstrak. Pada tahap screening, sebanyak 259 artikel dikeluarkan karena tidak sesuai dengan fokus penelitian, tidak membahas prediksi penjualan, tidak menggunakan data mining atau machine learning, tidak memakai algoritma klasifikasi atau regresi, tidak memuat evaluasi model, atau tidak memenuhi jenis publikasi ilmiah yang ditetapkan.

Sebanyak 87 artikel kemudian dinilai kelayakannya melalui pemeriksaan full-text. Pada tahap ini, 62 artikel dikeluarkan karena beberapa alasan. Sebanyak 15 artikel tidak menjelaskan metode data mining secara jelas, 13 artikel tidak berfokus pada prediksi penjualan, 10 artikel tidak menyajikan metrik evaluasi model, 9 artikel tidak menggunakan algoritma klasifikasi atau regresi, 5 artikel tidak tersedia dalam bentuk full-text, dan 10 artikel memakai data yang tidak relevan dengan penjualan. Setelah seluruh proses seleksi selesai, sebanyak 25 artikel memenuhi kriteria inklusi.



Gambar 1. Diagram PRISMA

Data dari 25 artikel inklusi diekstraksi menggunakan lembar ekstraksi yang disusun sesuai kebutuhan penelitian. Informasi yang diambil meliputi nama penulis dan tahun, negara atau sektor penelitian, jenis data penjualan, sumber data, jumlah data, algoritma yang digunakan, kategori algoritma, teknik praproses, metode validasi, metrik evaluasi, hasil utama, serta celah penelitian yang dinyatakan atau dapat ditarik dari artikel. Ekstraksi data dilakukan untuk memastikan setiap artikel dapat dibandingkan secara seimbang meskipun memiliki objek, algoritma, dan ukuran data yang berbeda.

Analisis SLR dilakukan dengan pendekatan deskriptif dan tematik. Analisis deskriptif digunakan untuk menghitung distribusi artikel menurut tahun, basis data, sektor penjualan, jenis algoritma, dan metrik evaluasi. Analisis tematik digunakan untuk mengelompokkan isi artikel ke dalam tema utama, yaitu penggunaan algoritma klasifikasi, penggunaan algoritma regresi, perbandingan performa algoritma, peran praproses data, pemilihan fitur penjualan, dan implikasi model prediksi bagi keputusan bisnis. Hasil sintesis tidak diarahkan untuk menentukan satu ukuran efek, melainkan untuk menunjukkan kecenderungan metode, keunggulan algoritma, kekurangan pendekatan terdahulu, serta kebutuhan riset berikutnya.

HASIL DAN PEMBAHASAN

Hasil penelitian ini disusun dari 25 artikel inklusi yang diperoleh melalui metode systematic literature review. Seluruh artikel yang dianalisis terbit pada tahun 2021–2025 dan membahas penerapan data mining atau machine learning untuk prediksi penjualan. Fokus kajian diarahkan pada penggunaan algoritma klasifikasi, algoritma regresi, model hibrida, serta pendekatan pendukung seperti association rule, RFM, dan analisis sentimen. Artikel yang dianalisis menunjukkan bahwa prediksi penjualan tidak hanya digunakan untuk memperkirakan jumlah atau nilai penjualan, tetapi juga untuk membantu pengambilan keputusan pada persediaan, promosi, rekomendasi produk, strategi harga, dan pengelolaan pelanggan.

Tabel 3. Distribusi Artikel Inklusi Berdasarkan Tahun Publikasi

Tahun Publikasi	Jumlah Artikel	Persentase
2021	8	32%
2022	5	20%
2023	5	20%
2024	6	24%
2025	1	4%
Total	**25**	**100%**

Tabel 3 menunjukkan bahwa artikel paling banyak terbit pada tahun 2021 dengan jumlah 8 artikel atau 32%. Publikasi tahun 2024 juga cukup tinggi yaitu 6 artikel atau 24%. Sebaran ini memperlihatkan bahwa prediksi penjualan berbasis data mining tetap menjadi perhatian penelitian setelah perubahan besar pada pola belanja digital. Artikel tahun 2025 masih berjumlah 1 artikel karena tahun publikasi tersebut belum selesai saat proses seleksi dilakukan. Sebaran tahun ini juga menunjukkan bahwa penelitian mengenai prediksi penjualan tidak berhenti pada satu periode tertentu, melainkan terus berkembang seiring meningkatnya kebutuhan bisnis terhadap keputusan berbasis data.

Tabel 4. Distribusi Artikel Berdasarkan Jenis Pendekatan Algoritma

Jenis Pendekatan Algoritma	Jumlah Artikel	Persentase	Sumber
Klasifikasi	4	16%	Ba'abbad et al. (2021); Permana et al. (2022); Johannes & Alamsyah (2021); Purwanto et al. (2025)
Regresi	16	64%	Zhang (2021); Merdas & Mousa (2023); Raizada & Saini (2021); Zubair et al. (2024); Ganguly & Mukherjee (2024); Li (2024)
Hibrida dan pendekatan pendukung	5	20%	Zhao & Keikhosrokiani (2022); Chalapathy & Josephine (2022); Poli & Sikder (2023); Ezeilo et al. (2022); Zhang et al. (2021)
Total	**25**	**100%**	

Tabel 4 memperlihatkan bahwa algoritma regresi menjadi pendekatan yang paling banyak digunakan. Sebanyak 16 artikel atau 64% memakai pendekatan regresi untuk memprediksi nilai penjualan, jumlah permintaan, omzet, atau volume produk. Klasifikasi digunakan pada 4 artikel atau 16%. Pendekatan ini lebih banyak dipakai untuk membedakan kelas penjualan, seperti produk laris dan tidak laris atau peluang pembelian pelanggan. Sebanyak 5 artikel atau 20% memakai pendekatan hibrida dan pendukung. Pola ini menunjukkan bahwa prediksi penjualan membutuhkan pendekatan yang berbeda sesuai dengan bentuk luaran yang ingin dicapai. Jika luaran berupa angka, regresi lebih banyak dipilih. Jika luaran berupa kategori, klasifikasi menjadi pilihan utama.

Tabel 5. Karakteristik Artikel Inklusi Berdasarkan Objek dan Data

No.	Penulis (Tahun)	Objek Penelitian	Data Penelitian	Metode	Hasil Penelitian
1	Ba'abbad et al. (2021)	Aplikasi bisnis umum	Data historis lintas bidang	Classification algorithms	Algoritma klasifikasi membantu memilih model prediksi yang sesuai untuk kebutuhan bisnis

No.	Penulis (Tahun)	Objek Penelitian	Data Penelitian	Metode	Hasil Penelitian
2	Zhang (2021)	Promosi e-commerce	Data promosi dan penjualan	Support Vector Regression	Model gabungan lebih akurat daripada SVM tunggal
3	Permana et al. (2022)	Penjualan produk	Data Kaggle 2013–2015	XGBoost, Decision Tree, Random Forest, Linear Regression, Naive Bayes	XGBoost menghasilkan RMSE training 0,68% dan testing 0,79%
4	Johannes & Alamsyah (2021)	UMKM footwear Tokopedia	Viewers, harga, jenis sepatu, jumlah terjual	Decision Tree	Jumlah produk terjual dapat diprediksi dari viewers, harga, dan jenis sepatu
5	Merdas & Mousa (2023)	Food sales	Dataset dengan korelasi fitur tinggi dan rendah	SVM, LASSO, Bagging, Gradient Boosting, Random Forest, Decision Tree	Algoritma terbaik berbeda sesuai korelasi fitur
6	Zhao & Keikhosrokiani (2022)	B2C dan rekomendasi produk	Historical sales dan perilaku pengguna	XGBoost, Random Forest, Apriori	XGBoost lebih baik dengan F1-score sekitar 0,789
7	Raizada & Saini (2021)	Walmart retail	Data 45 outlet Walmart	Multiple Linear Regression, Random Forest, KNN, SVM, Extra Tree	Prediksi memakai penjualan historis, hari libur, suhu, harga bahan bakar, CPI, dan tingkat pengangguran
8	Zubair et al. (2024)	Big Mart retail	Big Mart sales	Linear Regression, Random Forest, XGBoost	Random Forest Regression memperoleh R^2 tertinggi 0,545
9	Malik et al. (2024)	Rossmann retail	Rossmann sales	XGBoost, FB-Prophet, ARIMA	Model diuji dengan R^2, RMSE, dan MAE
10	Ganguly & Mukherjee (2024)	Retail sales	Dataset retail musiman	Random Forest, Gradient Boosting, SVR, XGBoost	Optimized Random Forest memperoleh R^2 0,945
11	Rao et al. (2021)	Retail sales	Training dataset retail	KNN, Multinomial, AdaBoost	Decision Tree dengan AdaBoost digunakan untuk memprediksi outlet sales
12	Bauskar (2022)	ERP retail	Walmart sales Kaggle	Random Forest Regression	Random Forest memperoleh MAE 19,79 dan R^2 0,94
13	Chalapathy & Josephine (2022)	E-commerce	Data produk, pelanggan,	RFM clustering dan regressor	Integrasi hasil cluster ke regressor meningkatkan prediksi

No.	Penulis (Tahun)	Objek Penelitian	Data Penelitian	Metode	Hasil Penelitian
			dan transaksi		
14	Deepa & Raghuram (2021)	Rossmann store	Data penjualan 3 tahun pada 1.115 toko	ARIMA, Seasonal Naive, Exponential Smoothing	ARIMA menjadi model terbaik dengan akurasi 74%
15	Panga (2023)	E-commerce trends	Historical sales	Linear Regression, Polynomial Regression, Random Forest, Gradient Boosting	Model hibrida diarahkan untuk meningkatkan akurasi prediksi
16	Poli & Sikder (2023)	Retail industry	Sales transaction	Apriori	Frequent itemset membantu membaca pola pembelian
17	Li (2024)	Online retail UK	E-commerce sales	Linear Regression, Decision Tree, Random Forest	Model dipilih melalui MAE, MSE, dan R^2
18	Ezeilo et al. (2022)	Omnichannel retail	Data retail lintas kanal	Hybrid machine learning	Model hibrida membantu menangani data retail lintas kanal
19	Shahar et al. (2024)	Online store	Data toko ritel	XGBoost dan RFE	XGBoost mencapai akurasi 98,23%
20	Krishna Madhav et al. (2023)	ERP sales	Public sales dataset	LSTM, SVM, XGBoost	LSTM memperoleh R^2 99,34%
21	Purwanto et al. (2025)	E-commerce Indonesia	Data transaksi e-commerce	Decision Tree, Random Forest, SVM, Naive Bayes, KNN	Random Forest mencapai akurasi 97,5% dan AUC 0,991
22	Ikram et al. (2023)	Big Mart sales	Big Mart dataset	SMO Regression, Linear Regression, MLP, Random Forest, M5P	Regresi dibandingkan melalui MAE, RAE, RMSE, dan RRSE
23	Lemos et al. (2021)	Harga logam mulia	Historical price data	Regression models	Model regresi digunakan untuk prediksi harga masa depan
24	Zhang et al. (2021)	Flash sales platform	Demand dan review sentiment	Sentiment analysis dan autoregressive model	Sentimen ulasan pelanggan meningkatkan akurasi forecast demand
25	Soltanineja d et al. (2024)	Mahram Food Industries	Data penjualan perusahaan	Neural network, regresi, time series	Neural network memperoleh MAD 28,33 dan MADP 10,2%

Keterangan: SVM = Support Vector Machine; LASSO = Least Absolute Shrinkage and Selection Operator; B2C = Business to Consumer; KNN = K-Nearest Neighbors; CPI = Consumer Price Index; ERP = Enterprise Resource Planning; RFM = Recency, Frequency, and Monetary; ARIMA = AutoRegressive Integrated Moving Average; UK = United Kingdom; RFE = Recursive Feature Elimination; LSTM = Long Short-Term Memory; SMO = Sequential Minimal Optimization; MLP = Multilayer Perceptron; M5P = Model Tree M5 Prime; MAE = Mean Absolute Error; MSE = Mean Squared Error; RMSE = Root Mean Squared Error; RAE = Relative Absolute Error; RRSE = Root Relative Squared Error; AUC = Area Under the Curve; MAD = Mean Absolute Deviation; MADP = Mean Absolute Deviation Percentage.

Tabel 5 menunjukkan bahwa objek penelitian didominasi oleh sektor retail dan e-commerce. Dua sektor ini banyak dikaji karena memiliki data transaksi yang besar, cepat berubah, dan dapat dihubungkan dengan promosi, harga, hari libur, perilaku pelanggan, serta kanal penjualan. Artikel yang memakai data retail umumnya memanfaatkan variabel penjualan historis, suhu, harga bahan bakar, indeks harga konsumen, dan tingkat pengangguran. Artikel yang memakai data e-commerce lebih banyak memanfaatkan perilaku pengguna, viewers, ulasan pelanggan, dan data transaksi digital. Keunggulan algoritma tidak sama pada setiap artikel. Random Forest unggul pada beberapa data retail dan e-commerce. XGBoost unggul pada prediksi berbasis fitur transaksi. LSTM unggul pada data penjualan yang memiliki pola waktu panjang.

Tabel 6. Frekuensi Penggunaan Algoritma pada Artikel Inklusi

Algoritma	Jumlah Kemunculan	Jenis Penggunaan	Kecenderungan Hasil
Random Forest/Random Forest Regression	10	Klasifikasi dan regresi	Sering unggul pada data tabular retail dan e-commerce
XGBoost	7	Klasifikasi dan regresi	Banyak digunakan karena akurasi tinggi pada data transaksi
Linear/Multiple/Polynomial Regression	7	Regresi	Sering menjadi pembanding awal dalam prediksi penjualan
Decision Tree	6	Klasifikasi dan regresi	Mudah dibaca sebagai aturan keputusan penjualan
SVM/SVR	6	Klasifikasi dan regresi	Digunakan pada data yang membutuhkan fungsi prediksi stabil
KNN	3	Klasifikasi dan regresi	Lebih sering dipakai sebagai pembanding
Naive Bayes	2	Klasifikasi	Digunakan untuk prediksi kelas penjualan
LSTM/Neural Network	2	Regresi dan time series	Kuat pada data berpola waktu panjang
Apriori/Association Rule	2	Pendekatan pendukung	Membantu membaca pola pembelian produk
ARIMA/Prophet/Exponential Smoothing	3	Time series	Digunakan untuk prediksi berbasis deret waktu

Tabel 6 memperlihatkan bahwa Random Forest menjadi algoritma yang paling sering muncul. Algoritma ini digunakan pada klasifikasi dan regresi karena mampu menangani banyak fitur dan pola hubungan yang tidak selalu linear. XGBoost juga banyak digunakan dan sering menunjukkan kinerja tinggi pada data transaksi. Linear Regression dan bentuk regresi lain tetap digunakan karena mudah dipahami dan dapat menjadi pembanding awal. Decision Tree juga sering muncul karena hasilnya lebih mudah dijelaskan kepada pengguna bisnis. Temuan ini menegaskan bahwa algoritma yang sering digunakan bukan hanya algoritma yang akurat, tetapi juga algoritma yang dapat diterjemahkan menjadi keputusan operasional.

Tabel 7. Metrik Evaluasi pada Artikel Inklusi

Kelompok Evaluasi	Metode Evaluasi	Jumlah Artikel	Fungsi Penilaian
Evaluasi klasifikasi	Accuracy, Precision, Recall, F1-score, AUC, Log Loss	6	Menilai ketepatan model dalam membedakan kelas penjualan
Evaluasi regresi	MAE, MSE, RMSE, MAPE, MADP, R^2	17	Menilai kesalahan prediksi dan kemampuan model menjelaskan variasi penjualan
Evaluasi deret waktu	Forecast accuracy, forecast error, ARIMA fit	3	Menilai ketepatan prediksi berbasis urutan waktu
Evaluasi pola pembelian	Support, Confidence	2	Menilai kekuatan hubungan antarproduk
Evaluasi kestabilan model	Cross-validation, train-test split, hyperparameter tuning	7	Menilai kestabilan model pada data latih dan data uji

Tabel 7 menunjukkan bahwa metrik regresi paling banyak digunakan. Hal ini sesuai dengan dominasi penelitian yang memprediksi nilai penjualan atau jumlah permintaan. MAE, MSE, RMSE, MAPE, MADP, dan R^2 digunakan untuk melihat jarak antara nilai aktual dan nilai prediksi. Pada artikel klasifikasi, accuracy, precision, recall, F1-score, AUC, dan Log Loss digunakan untuk melihat kemampuan model membedakan kelas penjualan. Perbedaan metrik ini penting karena setiap jenis luaran membutuhkan ukuran penilaian yang berbeda. Prediksi kelas penjualan membutuhkan metrik klasifikasi. Prediksi angka penjualan membutuhkan metrik regresi.

Tabel 8. Penelitian dengan Hasil Algoritma Paling Kuat

Penulis (Tahun)	Algoritma	Nilai Evaluasi	Hasil Penelitian
Purwanto et al. (2025)	Random Forest	Akurasi 97,5%; AUC 0,991	Random Forest paling unggul untuk prediksi penjualan e-commerce Indonesia dibandingkan Decision Tree, SVM, Naive Bayes, dan KNN
Bauskar (2022)	Random Forest Regression	MAE 19,79; R^2 0,94	Random Forest Regression efektif untuk prediksi penjualan pada sistem ERP berbasis dataset Walmart
Ganguly & Mukherjee (2024)	Optimized Random Forest	R^2 0,945; RMSLE 1,172	Optimized Random Forest mampu menangani data retail musiman dan banyak keluarga produk
Zubair et al. (2024)	Random Forest Regression	R^2 0,545	Random Forest Regression menghasilkan nilai R^2 tertinggi pada dataset Big Mart
Permana et al. (2022)	XGBoost	RMSE training 0,68%; RMSE testing 0,79%	XGBoost menjadi model terbaik untuk klasifikasi prediksi penjualan produk
Zhao & Keikhosrokiani (2022)	XGBoost	F1-score sekitar 0,789	XGBoost lebih baik pada prediksi pembelian pelanggan berbasis perilaku pengguna
Shahar et al. (2024)	XGBoost	Akurasi 98,23%	XGBoost menunjukkan performa tinggi pada prediksi penjualan online store setelah EDA dan seleksi fitur
Krishna Madhav et al. (2023)	LSTM	R^2 99,34%; MAE 11,64; RMSE 13,57	LSTM paling kuat untuk data penjualan ERP dengan pola waktu panjang

Deepa & Raghuram (2021)	ARIMA	Akurasi 74%	ARIMA menjadi model terbaik dibandingkan seasonal naive dan exponential smoothing
Soltaninejad et al. (2024)	Neural Network	MAD 28,33; MADP 10,2%; MSE 6452	Neural Network lebih baik dibandingkan regresi multivariabel dan time series pada perusahaan makanan

Tabel 8 memperlihatkan bahwa Random Forest, XGBoost, LSTM, ARIMA, dan Neural Network menjadi algoritma yang paling menonjol pada artikel inklusi. Random Forest kuat pada data tabular retail dan e-commerce. XGBoost kuat pada data transaksi yang sudah melewati seleksi fitur. LSTM lebih sesuai untuk data penjualan yang memiliki pola waktu panjang. ARIMA tetap relevan untuk prediksi deret waktu meskipun performanya tidak selalu mengungguli machine learning modern. Neural Network menunjukkan hasil baik pada kasus perusahaan makanan. Hasil ini menegaskan bahwa pemilihan algoritma perlu mengikuti bentuk data dan tujuan prediksi.

PEMBAHASAN

Hasil penelitian menunjukkan bahwa penerapan data mining dalam prediksi penjualan telah berkembang sebagai pendekatan penting untuk membantu perusahaan membaca pola transaksi, perilaku pelanggan, perubahan permintaan, serta kebutuhan persediaan. Dari 25 artikel inklusi, kajian paling banyak ditemukan pada sektor retail dan e-commerce karena kedua sektor tersebut memiliki data transaksi yang besar, cepat berubah, dan mudah dikaitkan dengan variabel promosi, harga, hari libur, kanal penjualan, serta karakter pelanggan. Prediksi penjualan tidak lagi hanya dipakai untuk memperkirakan angka penjualan masa depan, tetapi juga untuk mendukung keputusan operasional seperti pengaturan stok, penentuan promosi, rekomendasi produk, dan segmentasi pelanggan. Hal ini terlihat pada penelitian Zhang (2021) yang menunjukkan bahwa data promosi e-commerce dapat digunakan untuk membaca pengaruh aktivitas promosi terhadap volume penjualan, serta Raizada dan Saini (2021) yang menggunakan data 45 outlet Walmart dengan variabel penjualan historis, promosi, hari libur, suhu, harga bahan bakar, indeks harga konsumen, dan tingkat pengangguran. Temuan ini memperlihatkan bahwa prediksi penjualan yang baik tidak hanya bergantung pada data penjualan masa lalu, tetapi juga membutuhkan variabel pendukung yang menggambarkan kondisi pasar dan perilaku pelanggan.

Algoritma klasifikasi memiliki peran penting ketika tujuan prediksi diarahkan untuk menentukan kelas atau status penjualan. Klasifikasi dapat membantu perusahaan membedakan produk laris dan tidak laris, peluang pembelian pelanggan, atau kategori performa penjualan. Ba'abbad et al. (2021) menegaskan bahwa algoritma klasifikasi banyak digunakan dalam aplikasi data mining karena dapat membantu analis memilih model prediksi yang sesuai dengan kebutuhan bisnis. Pada sektor e-commerce Indonesia, Purwanto et al. (2025) menunjukkan bahwa Random Forest memperoleh performa terbaik dibandingkan Decision Tree, Support Vector Machine, Naive Bayes, dan K-Nearest Neighbors dengan akurasi 97,5% dan AUC 0,991. Permana et al. (2022) juga menemukan bahwa XGBoost menjadi model terbaik dalam klasifikasi prediksi penjualan dengan RMSE training 0,68% dan RMSE testing 0,79%. Johannes dan Alamsyah (2021) memakai Decision Tree untuk memprediksi penjualan produk footwear pada Tokopedia melalui variabel viewers, harga, dan jenis sepatu. Kumpulan temuan tersebut memperlihatkan bahwa klasifikasi sangat berguna saat hasil prediksi perlu diterjemahkan menjadi keputusan yang sederhana, misalnya produk yang perlu diprioritaskan, produk yang perlu diberi promosi, atau produk yang membutuhkan pengendalian stok lebih cepat.

Pendekatan regresi menjadi pendekatan yang paling dominan karena prediksi penjualan umumnya membutuhkan luaran berbentuk angka, seperti jumlah permintaan, nilai omzet, volume produk, atau sales outlet. Random Forest Regression, XGBoost, Support Vector Regression, Linear Regression, Gradient Boosting, LSTM, ARIMA, dan Neural Network banyak digunakan karena mampu membaca hubungan antara penjualan dan variabel pendukung. Ganguly dan Mukherjee (2024) menunjukkan bahwa Optimized Random Forest memperoleh R^2 sebesar 0,945 dan mengungguli Linear Regression, Gradient Boosting, SVR, serta XGBoost pada data retail dengan pola

musiman dan banyak keluarga produk. Bauskar (2022) juga menemukan bahwa Random Forest Regression pada data Walmart menghasilkan MAE 19,79 dan R^2 0,94. Pada data ERP, Krishna Madhav et al. (2023) menunjukkan bahwa LSTM memperoleh R^2 99,34%, MAE 11,64, dan RMSE 13,57 sehingga lebih kuat dalam membaca pola waktu panjang dibandingkan SVM dan XGBoost. Shahar et al. (2024) juga memperlihatkan bahwa XGBoost mampu mencapai akurasi 98,23% setelah melalui Exploratory Data Analysis dan Recursive Feature Elimination. Temuan ini menegaskan bahwa regresi dan model sekuensial lebih tepat digunakan saat penelitian ingin memperkirakan nilai penjualan secara numerik, terutama pada data yang memiliki pola waktu, promosi, musim, dan variasi outlet.

Hasil penelitian memperlihatkan bahwa tidak ada satu algoritma yang selalu unggul untuk seluruh jenis data penjualan. Random Forest cenderung kuat pada data tabular retail dan e-commerce. XGBoost menonjol pada data transaksi yang telah melalui seleksi fitur. LSTM lebih sesuai untuk data penjualan yang memiliki urutan waktu panjang. Model hibrida juga mulai banyak digunakan karena prediksi penjualan sering membutuhkan gabungan antara segmentasi pelanggan, pola pembelian, rekomendasi produk, dan estimasi angka penjualan. Zhao dan Keikhosrokiani (2022) menggabungkan RFM, XGBoost, Random Forest, dan Apriori untuk prediksi pembelian serta rekomendasi produk. Chalapathy dan Josephine (2022) mengintegrasikan RFM clustering dengan regressor untuk meningkatkan prediksi penjualan e-commerce. Zhang et al. (2021) menambahkan faktor sentimen ulasan pelanggan ke model autoregressive pada flash sales platform. Ezeilo et al. (2022) juga menekankan pentingnya model hibrida pada retail omnichannel karena data penjualan berasal dari banyak kanal dan memiliki karakter yang beragam. Celah yang masih terlihat adalah terbatasnya riset yang menggabungkan klasifikasi dan regresi dalam satu alur prediksi. Penelitian berikutnya dapat mengembangkan model dua tahap, yaitu klasifikasi status produk atau pelanggan terlebih dahulu, lalu dilanjutkan dengan regresi untuk memperkirakan nilai penjualan pada setiap kelompok yang terbentuk. Pendekatan ini akan membuat hasil prediksi lebih mudah dipakai dalam keputusan bisnis, terutama bagi UMKM, retail lokal, dan toko online di Indonesia.

KESIMPULAN

Hasil penelitian menunjukkan bahwa penerapan data mining dalam prediksi penjualan telah berkembang sebagai pendekatan penting untuk membantu pelaku usaha membaca pola transaksi dan perilaku pelanggan secara lebih terukur. Dari 25 artikel inklusi yang dianalisis, pendekatan regresi menjadi kelompok algoritma paling dominan karena banyak penelitian berfokus pada prediksi nilai penjualan, jumlah permintaan, omzet, dan volume produk. Random Forest Regression, XGBoost, LSTM, ARIMA, dan Neural Network menjadi algoritma yang banyak digunakan untuk kebutuhan tersebut. Pendekatan klasifikasi juga memiliki peran penting karena mampu membantu perusahaan membedakan kelas penjualan, seperti produk laris dan tidak laris, peluang pembelian, atau status performa produk. Random Forest dan XGBoost menunjukkan hasil yang kuat pada beberapa penelitian karena mampu mengolah data transaksi dengan banyak fitur. Model hibrida juga mulai berkembang karena prediksi penjualan sering memerlukan gabungan antara segmentasi pelanggan, pola pembelian, rekomendasi produk, dan estimasi angka penjualan.

Implikasi penelitian ini menunjukkan bahwa prediksi penjualan berbasis data mining dapat membantu perusahaan dalam menyusun strategi persediaan, promosi, harga, distribusi, dan pengelolaan pelanggan. Pemilihan algoritma tidak dapat dilakukan hanya berdasarkan popularitas model. Pemilihan perlu menyesuaikan tujuan prediksi, bentuk data, ukuran data, kualitas fitur, dan kebutuhan keputusan bisnis. Untuk penelitian berikutnya, penggunaan data lokal dari UMKM, retail kecil, toko online, dan perusahaan Indonesia perlu diperluas agar hasil penelitian lebih dekat dengan kebutuhan bisnis nyata. Penelitian selanjutnya juga disarankan mengembangkan model dua tahap yang menggabungkan klasifikasi dan regresi. Tahap pertama digunakan untuk mengelompokkan status produk atau pelanggan. Tahap kedua digunakan untuk memperkirakan nilai penjualan pada setiap kelompok tersebut. Pendekatan ini dapat membuat hasil prediksi lebih mudah dipahami dan lebih bermanfaat bagi pengambilan keputusan.

REFERENSI

Ba'abbad, I., Althubiti, T., Alharbi, A., Alfarsi, K., & Rasheed, S. (2021). A short review of classification algorithms accuracy for data prediction in data mining applications. *Journal of Data Analysis and Information Processing, 9*(3), 162–174. https://doi.org/10.4236/jdaip.2021.93011

Bank Indonesia. (2025). *FEKDI x IFSE 2025: Digitalisasi ekonomi keuangan kunci penguatan ekonomi nasional*. Bank Indonesia.

Bauskar, S. (2022). Predictive analytics for sales forecasting in enterprise resource planning (ERP) systems using machine learning technique. *SSRN Electronic Journal*. https://doi.org/10.56726/IRJMETS26271

Chalapathy, N., & Josephine VL, H. (2022). Sales prediction scheme using RFM based clustering and regressor model for ecommerce company. In *Proceedings of the 4th International Conference on Information Management & Machine Intelligence* (pp. 1–6). Association for Computing Machinery. https://doi.org/10.1145/3590837.3590937

Deepa, K., & Raghuram, G. (2021). Sales forecasting using machine learning models. *Annals of the Romanian Society for Cell Biology, 25*(5), 3928–3936.

Douaioui, K., Oucheikh, R., Benmoussa, O., & Mabrouki, C. (2024). Machine learning and deep learning models for demand forecasting in supply chain management: A critical review. *Applied System Innovation, 7*(5), 93. DOI: 10.3390/asi7050093

Eglite, L., & Birzniece, I. (2022). Retail sales forecasting using deep learning: Systematic literature review. *Complex Systems Informatics and Modeling Quarterly, 30*, 53–62. DOI: 10.7250/csimq.2022-30.03

Elalem, Y. K., Maier, S., & Seifert, R. W. (2023). A machine learning-based framework for forecasting sales of new products with short life cycles using deep neural networks. *International Journal of Forecasting, 39*(4), 1874–1894. DOI: 10.1016/j.ijforecast.2022.09.005

Ensafi, Y., Amin, S. H., Zhang, G., & Shah, B. (2022). Time-series forecasting of seasonal items sales using machine learning: A comparative analysis. *International Journal of Information Management Data Insights, 2*(1), 100058. DOI: 10.1016/j.jjimei.2022.100058

Ezeilo, O. J., Ikponmwoba, S. O., Chima, O. K., Ojonugwa, B. M., & Adesuyi, M. O. (2022). Hybrid machine learning models for retail sales forecasting across omnichannel platforms. *Shodhshauryam, 5*(2), 175–190.

Ganguly, P., & Mukherjee, I. (2024). Enhancing retail sales forecasting with optimized machine learning models. In *2024 4th International Conference on Sustainable Expert Systems (ICSES)* (pp. 884–889). IEEE. https://doi.org/10.1109/ICSES63445.2024.10762950

Haselbeck, F., Killinger, J., Menrad, K., Hannus, T., & Grimm, D. G. (2022). Machine learning outperforms classical forecasting on horticultural sales predictions. *Machine Learning with Applications, 7*, 100239. DOI: 10.1016/j.mlwa.2021.100239

Hwang, S., Lee, Y., Jeon, B.-K., & Oh, S. H. (2025). Sales forecasting for new products using homogeneity-based clustering and ensemble method. *Electronics, 14*(3), 520. DOI: 10.3390/electronics14030520

Hwang, S., Yoon, G., Baek, E., & Jeon, B.-K. (2023). A sales forecasting model for new-released and short-term product: A case study of mobile phones. *Electronics, 12*(15), 3256. DOI: 10.3390/electronics12153256

Ikram, A., Butt, M. A., & Tariq, I. (2023). Comparative analysis of regression algorithms used to predict the sales of big marts. *Journal of Innovative Computing and Emerging Technologies, 3*(1). https://doi.org/10.56536/jicet.v3i1.53

Johannes, R., & Alamsyah, A. (2021). Sales prediction model using classification decision tree approach for small medium enterprise based on Indonesian e-commerce data. *arXiv*. https://doi.org/10.48550/arXiv.2103.03117

Kementerian Perdagangan Republik Indonesia. (2026). *Kinerja perdagangan melalui sistem elektronik (PMSE) di Indonesia tahun 2024*. Badan Kebijakan Perdagangan.

Krishna Madhav, J., Varun, B., Niharika, K., Srinivasa Rao, M., & Laxmana Murthy, K. (2023). Optimising sales forecasts in ERP systems using machine learning and predictive analytics.

Journal of Contemporary Education Theory & Artificial Intelligence. https://doi.org/10.47991/2023/JCETAI-104

Lemos, N., Pawaskar, I., Ramchandani, D., & Poojary, T. (2021). Intelligent sales prediction using machine learning. *International Journal of Business Forecasting and Marketing Intelligence, 34*(3), 234–247.

Li, X. (2024). Predicting e-commerce sales with three machine learning models. In *Proceedings of the 2024 2nd International Conference on Management Innovation and Economy Development (MIED 2024)* (pp. 445–454). Atlantis Press. https://doi.org/10.2991/978-94-6463-542-3_52

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2022). The M5 competition: Background, organization, and implementation. *International Journal of Forecasting, 38*(4), 1325–1336. DOI: 10.1016/j.ijforecast.2021.07.007

Malik, S., Khan, M., Abid, M. K., & Aslam, N. (2024). Sales forecasting using machine learning algorithm in the retail sector. *Journal of Computing & Biomedical Informatics, 6*(2), 282–294.

Massaro, A., Maritati, V., Galiano, A., & Birardi, V. (2021). Augmented data and XGBoost improvement for sales forecasting in large scale retail. *Applied Sciences, 11*(17), 7793. DOI: 10.3390/app11177793

Merdas, H. M., & Mousa, A. H. (2023). Food sales prediction model using machine learning techniques. *International Journal of Electrical and Computer Engineering, 13*(6), 6578–6585. https://doi.org/10.11591/ijece.v13i6.pp6578-6585

Nasseri, M., Falatouri, T., Brandtner, P., & Darbanian, F. (2023). Applying machine learning in retail demand prediction: A comparison of tree-based ensembles and long short-term memory-based deep learning. *Applied Sciences, 13*(19), 11112. DOI: 10.3390/app131911112

Nurdin, N., Susanti, E., Aidilof, H. A. K., & Priyanto, D. (2022). Comparison of Naive Bayes and Dempster Shafer methods in expert system for early diagnosis of COVID-19. *MATRIK: Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer, 22*(1), 215–228. DOI: 10.30812/matrik.v22i1.2280

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... & Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *bmj*, *372*.

Panarese, G., Gallo, M., & Cesarelli, G. (2022). Developing and preliminary testing of a machine learning-based platform for sales forecasting using a gradient boosting approach. *Applied Sciences, 12*(21), 11054. DOI: 10.3390/app122111054

Panga, N. K. R. (2023). Forecasting e-commerce trends: Utilizing linear regression, polynomial regression, random forest, and gradient boosting for accurate sales and demand prediction. *International Journal of HRM and Organizational Behavior, 11*(3), 11–26.

Permana, S., Rosadi, R., & Nikki, N. (2022). Application of classification algorithm for sales prediction. *TEKNOKOM, 5*(2), 119–124. https://doi.org/10.31943/teknokom.v5i2.77

Poli, N. S., & Sikder, A. S. (2023). Predictive analysis of sales using the Apriori algorithm: A comprehensive study on sales forecasting and business strategies in the retail industry. *International Journal of Imminent Science & Technology, 1*(1), 1–16. https://doi.org/10.70774/ijist.v1i1.1

Prabu, C. R., Ravindran, R., Varma, K. S. A., Sri Rohith, K. S., Sowmya, G., & Remya, M. S. (2025). Comprehensive review on sales prediction models. *Procedia Computer Science, 259*, 1218–1227. DOI: 10.1016/j.procs.2025.04.077

Purwanto, E., Cipto Utomo, B. P., Permatasari, H., & Mohd, F. (2026). Comparative analysis of classification models for sales prediction in e-commerce: Decision Tree, Random Forest, SVM, Naive Bayes, and KNN. *Jurnal Teknik Informatika (JUTIF), 6*(6), 5899–5915. DOI: 10.52436/1.jutif.2025.6.6.5224

Purwanto, E., Utomo, B. P. C., Permatasari, H., & Mohd, F. (2025). Comparative analysis of classification models for sales prediction in e-commerce: Decision Tree, Random Forest, SVM, Naive Bayes, and KNN. *Jurnal Teknik Informatika (JUTIF), 6*(6), 5899–5915. https://doi.org/10.52436/1.jutif.2025.6.6.5224

Raizada, S., & Saini, J. R. (2021). Comparative analysis of supervised machine learning techniques for sales forecasting. *International Journal of Advanced Computer Science and Applications, 12*(11), 102–110. https://doi.org/10.14569/IJACSA.2021.0121112

Rao, B. S., Sujatha, K., Rao, N. S., & Kumar, T. N. (2021). Retail sales prediction using machine learning algorithms. *Turkish Online Journal of Qualitative Inquiry, 12*(1), 315–322.

Salamah, S., Abdullah, D., & Nurdin, N. (2024). Comparative analysis of K-Means and K-Medoids to determine study programs. *International Journal of Engineering, Science and Information Technology, 5*(1), 167–176. DOI: 10.52088/ijesty.v5i1.673

Salsabila, T., Nurdin, N., & Retno, S. (2025). Comparison of K-Medoids and K-Means result for regional clustering of capture fisheries in Aceh Province. *International Journal of Engineering, Science and Information Technology, 5*(2), 282–289. DOI: 10.52088/ijesty.v5i2.829

Saptadi, S., Budiawan, W., Fadhila, R. N., Sumbodo, R. K., & Saputra, D. W. N. (2026). Machine learning applied to sales prediction modeling: A systematic literature review. *Multidisciplinary Reviews, 9*(3), e2026127. DOI: 10.31893/multirev.2026127

Shahar, N. A. M., Mutalib, S., Halim, S. A., & Ramdhan, W. (2024). Sales forecasting using machine learning methods for online store. In *International Conference on Innovation & Entrepreneurship in Computing, Engineering & Science Education (InvENT 2024)* (pp. 41–52). Atlantis Press. https://doi.org/10.2991/978-94-6463-589-8_6

Soltaninejad, M., Aghazadeh, R., Shaghaghi, S., & Zarei, M. (2024). Using machine learning techniques to forecast Mehram company's sales: A case study. *Journal of Business and Management Studies, 6*(2), 42–53. https://doi.org/10.32996/jbms.2024.6.2.4

Tang, Z., & Dong, S. (2021). A total sales forecasting method for a new short life-cycle product in the pre-market period based on an improved evidence theory: Application to the film industry. *International Journal of Production Research, 59*(22), 6776–6790. DOI: 10.1080/00207543.2020.1825861

Zhang, M., Wang, Y., & Wu, Z. (2021). Data mining algorithm for demand forecast analysis on flash sales platform. *Complexity, 2021*, Article 6648009. https://doi.org/10.1155/2021/6648009

Zhang, Y. (2021). Sales forecasting of promotion activities based on the cross-industry standard process for data mining of e-commerce promotional information and support vector regression. *Journal of Computers, 32*(1), 212–225. https://doi.org/10.3966/199115992021023201018

Zhao, X., & Keikhosrokiani, P. (2022). Sales prediction and product recommendation model through user behavior analytics. *Computers, Materials & Continua, 70*(2). https://doi.org/10.32604/cmc.2022.019750

Zubair, M., Waleed, A., Rehman, A., Ahmad, F., Islam, M., & Javed, S. (2024). Machine learning insights into retail sales prediction: A comparative analysis of algorithms. In *2024 Horizons of Information Technology and Engineering (HITE)* (pp. 1–7). IEEE. https://doi.org/10.1109/HITE63532.2024.10777132.