JURNAL ILMIAH FIFO P-ISSN 2085-4315
Vol. 18 No. 1 (2026) 13-21 E-ISSN 2502-8332
DOI: 10.22441/fifo.2026.v18i1.002

Predicting Consumer Purchasing Behavior Using Random Forest on Retail Transaction Data

Ningsiah[1], Nur Aminudin[2*]

[1]Department of Pharmacy, Faculty of Health, Aisyah University, Indonesia
[2]Department of Software Engineering, Faculty of Technology and Informatics, Aisyah University, Indonesia
Email: [1]ningsiah@aisyahuniversity.ac.id, [2]nuraminudin@aisyahuniversity.ac.id
Corresponding Author[*]

(***received*****:** 13-02-26, ***revised*****:** 21-02-26, ***accepted*****:** 02-03-26)

Abstrak

Transformasi digital yang pesat di sektor ritel telah menghasilkan volume data transaksi konsumen dalam jumlah besar yang tersimpan dalam sistem informasi ritel. Meskipun data tersebut memiliki nilai strategis dalam mendukung pengambilan keputusan, pemanfaatannya masih sering terbatas pada pelaporan deskriptif. Penelitian ini bertujuan untuk menganalisis dan memprediksi perilaku pembelian konsumen melalui integrasi analitik prediktif berbasis machine learning ke dalam sistem informasi ritel dengan menggunakan dataset transaksi ritel dari Kaggle. Metodologi penelitian meliputi prapemrosesan data, analisis data eksploratif, seleksi fitur, serta pengembangan model prediktif menggunakan algoritma regresi logistik, decision tree, dan random forest. Kinerja model dievaluasi menggunakan metrik accuracy, precision, recall, dan ROC–AUC. Hasil penelitian menunjukkan bahwa model random forest mengungguli algoritma lainnya dengan tingkat accuracy sebesar 88,76%, precision sebesar 87,92%, dan recall sebesar 86,48%, yang mencerminkan kemampuan diskriminasi yang lebih unggul. Temuan ini menegaskan bahwa metode pembelajaran berbasis ensemble efektif dalam menangkap pola pembelian konsumen yang kompleks dan non-linear. Secara teoretis, penelitian ini berkontribusi dengan memperluas peran sistem informasi ritel dari sekadar alat pelaporan deskriptif menjadi sistem pendukung keputusan prediktif. Secara praktis, penelitian ini menyediakan kerangka analitis yang kuat untuk mendukung optimalisasi persediaan, strategi promosi yang lebih terarah, serta personalisasi layanan dalam lingkungan ritel berbasis data.

Kata Kunci: Analitik Prediktif, Data Transaksi, Machine Learning, Perilaku Belanja Konsumen, Sistem Informasi Ritel.

Abstract

The rapid digital transformation in the retail sector has generated massive volumes of consumer transaction data stored within retail information systems. Although these data hold strategic value for decision-making, their utilization often remains limited to descriptive reporting. This study aims to analyze and predict consumer purchasing behavior by integrating machine learning–based predictive analytics into retail information systems using the Kaggle retail transaction dataset. The research methodology includes data preprocessing, exploratory data analysis, feature selection, and predictive model development using logistic regression, decision tree, and random forest algorithms. Model performance was evaluated using accuracy, precision, recall, and ROC–AUC metrics. The results indicate that the random forest model outperformed the other algorithms, achieving an accuracy of 88.76%, precision of 87.92%, and recall of 86.48%, demonstrating superior discriminative capability. These findings confirm that ensemble-based learning methods effectively capture complex and non-linear consumer purchasing patterns. The study contributes theoretically by extending the role of retail information systems from descriptive reporting tools to predictive decision-support systems, while practically providing a robust analytical framework to support inventory optimization, targeted promotion strategies, and personalized service delivery in data-driven retail environments.

Keywords*: Predictive Analytics, Transaction Data, Machine Learning, Consumer Purchasing Behavior, Retail Information Systems.*

1. INTRODUCTION

Digital transformation has fundamentally reshaped the way retail organizations manage data and formulate strategic decisions. in information technology, supported by the availability of big data and increasingly integrated information systems, have enabled retail firms to collect, store, and process large-scale consumer transaction data in near real time [1], [2]. Such transaction data do not merely record purchase histories, but also reflect consumer preferences, purchasing habits, and dynamic behavioral patterns. In this context, retail information systems serve as a strategic infrastructure that integrates data resources, business processes, and analytical capabilities to support more accurate and data-driven decision making [3], [4].

Nevertheless, the growing volume, variety, and velocity of transaction data present significant analytical challenges. Conventional analytical approaches that focus primarily on descriptive and retrospective analysis are increasingly insufficient to extract strategic value from complex consumer data [5], [6]. This condition has driven the adoption of more advanced analytical approaches, particularly predictive analytics, which combines statistical techniques, data mining, and machine learning to model and forecast consumer purchasing behavior based on historical patterns [7], [8], [9]. When effectively integrated into retail information systems, predictive analytics has the potential to enhance consumer behavior understanding, optimize marketing strategies, and strengthen organizational competitiveness in an increasingly data-intensive retail environment.

Despite these advancements, the utilization of consumer transaction data in many retail organizations remains largely confined to operational reporting and historical analysis [8], [10], [11]. Existing retail information systems often lack the integration of advanced analytical capabilities required to comprehensively understand and predict consumer purchasing behavior. As a result, transaction data with substantial strategic potential are underutilized in supporting proactive managerial decisions, such as inventory planning, promotional strategy formulation, and service personalization [12], [13], [14]. This situation indicates a clear gap between the availability of transaction data within retail information systems and the analytical capacity needed to generate high-value insights for decision support.

The urgency of addressing this issue has increased alongside rapid changes in consumer behavior driven by digital technologies and intensified competition in the retail sector. Without integrated predictive analytical support, organizations risk relying on intuition-based decisions or purely descriptive analyses that are insufficiently adaptive to evolving purchasing patterns. Therefore, there is a growing need for an information system–based approach that extends beyond data storage and reporting functions to transform historical transaction data into predictive models capable of anticipating future consumer behavior trends.

Previous studies have explored consumer purchasing behavior analysis using data mining and machine learning techniques [7], [15], [16]. However, much of the existing research focuses on algorithmic performance comparisons or improvements in predictive accuracy, often without explicitly situating these methods within the functional role of integrated retail information systems. Furthermore, many studies emphasize static descriptive analysis or consumer segmentation, thereby underexploring the potential of predictive analytics to capture the evolving dynamics of consumer purchasing behavior. Another limitation lies in the relatively limited use of actual retail transaction data within a comprehensive information system framework, which constrains the practical applicability and generalizability of prior findings [12], [17], [18], [19].

Previous studies have primarily focused on improving predictive accuracy through algorithmic comparisons or customer segmentation techniques without explicitly embedding predictive models within the architectural and functional context of retail information systems [15], [20], [21]. While several studies demonstrated the effectiveness of machine learning in forecasting consumer behavior, limited attention has been given to how predictive analytics can be systematically integrated into retail information system infrastructures to enhance strategic decision-making capabilities. Moreover, empirical investigations that utilize transaction-level retail datasets within a structured information system framework remain relatively scarce. This gap highlights the need for research that bridges machine learning modeling with information systems integration to generate both analytical rigor and practical business value.

In response to these limitations, this study aims to analyze and predict consumer purchasing behavior through the integration of predictive analytics within retail information systems. Specifically, this research develops a transaction data–driven analytical framework to identify purchasing behavior patterns and construct predictive models that support more proactive and data-driven decision making. The expected contributions of this study include a theoretical contribution by extending information systems research through the integration of predictive analytics in consumer behavior analysis, as well as a practical contribution by providing empirical evidence and analytical models that can be utilized by retail information system managers to enhance business strategy effectiveness, operational planning, and service personalization.

This study contributes to the literature in three significant ways. First, it extends information systems research by conceptualizing predictive analytics as an embedded capability within retail information system architectures rather than as a standalone analytical tool. Second, it provides methodological contributions through a structured predictive modeling framework incorporating feature selection, cross-validation, and multi-metric evaluation to ensure robustness and replicability. Third, from a practical perspective, the study

JURNAL ILMIAH FIFO P-ISSN 2085-4315
Vol. 18 No. 1 (2026) 13-21 E-ISSN 2502-8332
DOI: 10.22441/fifo.2026.v18i1.002

offers empirical evidence demonstrating how random forest–based predictive models can enhance retail decision-making processes, particularly in inventory planning, promotion targeting, and service personalization.

2. RESEARCH METHOD

This study employs a quantitative approach with a predictive analytics design based on machine learning to analyze consumer purchasing behavior within the context of retail information systems. The methodological framework is structured into five main stages: (1) data acquisition, (2) data preprocessing and exploration, (3) feature selection, (4) predictive model development, and (5) model performance evaluation. This structured workflow ensures methodological rigor, replicability, and the potential integration of the developed models into retail information system architectures [22], [23], [24], [25].

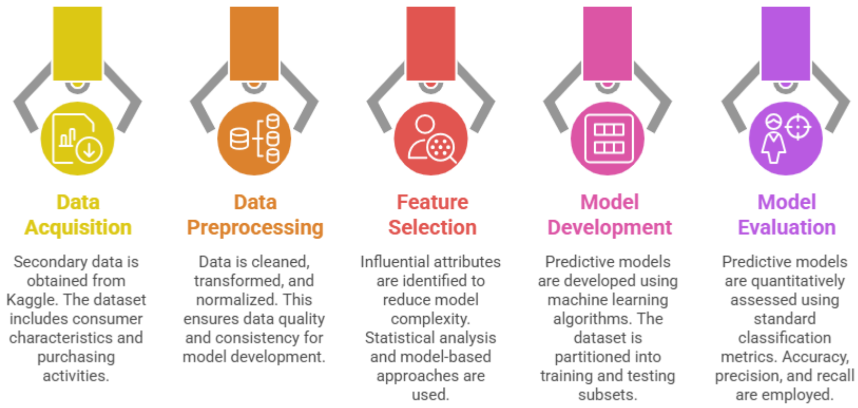

Figure 1. Research Method Stages

2.1. Data Source and Characteristics

The data used in this study consist of secondary data obtained from the Kaggle platform in the form of a retail transaction dataset. The dataset includes attributes that represent consumer characteristics and purchasing activities, such as demographic variables, product categories, transaction frequency, expenditure value, and other behavioral indicators. The dataset was selected due to its open accessibility, well-documented structure, and frequent use in consumer behavior analytics research, which facilitates validation and comparative analysis [26], [27].

The unit of analysis is individual consumers, with each record representing a documented transaction within a specific observation period. The dependent variable is defined based on the purchasing behavior indicator to be predicted, while the remaining attributes are treated as predictor variables.

2.2. Data Preprocessing and Exploratory Analysis

Data preprocessing is conducted to ensure data quality and consistency prior to model development [28], [29], [30]. This stage includes:

a. Data cleaning, involving the removal of duplicate records and the handling of missing values using appropriate imputation techniques.
b. Data transformation, particularly the conversion of categorical variables into numerical representations through encoding techniques.
c. Normalization or standardization, applied to numerical variables to ensure comparable scales and to prevent bias during model training.

Subsequently, exploratory data analysis (EDA) is performed to examine variable distributions, identify initial relationships among attributes, and detect underlying patterns related to consumer purchasing behavior. This step also serves to identify potential class imbalance issues in the target variable.

2.3. Feature Selection

Feature selection is applied to identify the most influential attributes contributing to the prediction of consumer purchasing behavior and to reduce model complexity. This process combines statistical analysis (e.g., correlation analysis and significance testing) with model-based approaches that assess feature importance. Attributes with minimal contribution are excluded to improve computational efficiency and to mitigate the risk of overfitting. The selected feature set forms the basis for subsequent predictive modeling [23], [31], [20].

2.4. Predictive Model Development

Predictive models are developed using three widely adopted machine learning algorithms in consumer behavior analysis [32], [33], [21]:

a. Logistic Regression, to model the probabilistic relationship between predictor variables and the target class.
b. Decision Tree, to capture non-linear patterns through rule-based data partitioning.
c. Random Forest, an ensemble method that aggregates multiple decision trees to enhance prediction accuracy and model robustness.

Random forest was selected due to its capability to handle high-dimensional transaction data, capture non-linear relationships, and mitigate overfitting through ensemble aggregation. In consumer purchasing behavior contexts characterized by heterogeneous and dynamic patterns, ensemble learning methods provide improved bias–variance tradeoff compared to single-model approaches. Prior studies in retail analytics have also demonstrated the robustness of random forest in modeling complex behavioral interactions [32], [21].

The dataset is partitioned into training and testing subsets using a predefined ratio (e.g., 80:20) to evaluate the generalization capability of the models. To enhance the reliability of model performance estimation, cross-validation is applied during the training phase.

2.5. Model Performance Evaluation

To quantitatively assess the performance of the predictive models, this study employs three standard classification evaluation metrics, namely accuracy, precision, and recall.

Accuracy, precision, and recall were selected to provide a comprehensive evaluation of classification performance. While accuracy reflects overall prediction correctness, precision assesses the reliability of positive predictions, and recall measures the model's sensitivity in identifying actual purchasing behavior patterns. The inclusion of ROC–AUC further strengthens performance evaluation by examining discriminative ability across threshold variations, ensuring model robustness in imbalanced retail datasets.

These metrics are derived from the confusion matrix, which consists of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) [34], [35], [36], [37], [38].

Accuracy measures the overall correctness of the model by comparing the number of correctly classified instances to the total number of observations. The accuracy metric is formulated as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Precision evaluates the proportion of correctly predicted positive instances relative to all instances predicted as positive. This metric reflects the reliability of positive class predictions and is defined as:

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

Recall, also referred to as sensitivity, measures the ability of the model to correctly identify all actual positive instances. Recall is particularly important in understanding how well the model captures relevant consumer purchasing behavior patterns. The recall metric is expressed as:

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

The simultaneous use of accuracy, precision, and recall provides a comprehensive evaluation of model performance, especially in cases where the class distribution is imbalanced. Models are compared based on these metrics, and the optimal model is selected by considering both predictive effectiveness and stability across the testing dataset.

3. RESULTS AND DISCUSSION

3.1. Data Preprocessing and Descriptive Analysis

The data preprocessing stage constitutes a critical initial step in this study to ensure data quality prior to predictive modeling. The consumer purchasing behavior dataset obtained from Kaggle was first subjected to a data cleaning process, which involved the removal of duplicate records and the treatment of missing values across several attributes. Missing values were handled using imputation techniques tailored to the characteristics of each variable to preserve the overall data distribution. In addition, categorical attributes were transformed into

P-ISSN 2085-4315
E-ISSN 2502-8332

numerical representations using appropriate encoding techniques to enable processing by machine learning algorithms.

Based on the results of descriptive statistical analysis, the preprocessed dataset exhibits characteristics that are representative for consumer purchasing behavior analysis. Descriptive statistics include minimum, maximum, mean, and standard deviation values for key numerical variables, such as purchase frequency and consumer expenditure. The analysis reveals substantial variation in expenditure-related variables, indicating heterogeneity in consumer purchasing patterns. This variation provides an essential foundation for predictive modeling, as it reflects diverse consumer behavior within retail information systems.

The distribution of consumer purchasing behavior was further examined to understand purchasing tendencies prior to predictive modeling. The results indicate that the majority of consumers fall within the medium purchasing behavior category, while the proportion of consumers with high purchasing frequency is relatively smaller. This partially imbalanced distribution highlights the importance of selecting appropriate evaluation metrics during the modeling stage, particularly precision and recall, to prevent model bias toward the majority class. These findings also suggest that retail information systems must accommodate diverse behavioral segments to effectively support data-driven decision making.

Overall, the preprocessing results confirm that the dataset is in a suitable and optimal condition for predictive analytics modeling. Descriptive statistical analysis and behavioral distribution assessment provide valuable preliminary insights into data characteristics and potential modeling challenges. These findings serve as a methodological foundation for subsequent machine learning model development and ensure that predictive outcomes can be interpreted validly within the context of retail information systems.

3.2. Model Performance Evaluation

The predictive analytics modeling stage was conducted to evaluate the capability of machine learning algorithms in predicting consumer purchasing behavior based on the preprocessed dataset. Three algorithms logistic regression, decision tree, and random forest were applied to the same dataset to ensure a fair and consistent comparison of model performance. The dataset was divided into training and testing subsets to assess the generalization ability of each model on previously unseen data.

Table 1 presents the performance evaluation results of the predictive analytics models using accuracy, precision, and recall metrics.

Table 1. Performance Evaluation of Predictive Analytics Models

model machine learning	accuracy (%)	precision (%)	recall (%)
Logistic Regression	82.45	80.12	78.96
Decision Tree	84.31	83.27	81.54
Random Forest	88.76	87.92	86.48

The evaluation results demonstrate that the random forest model outperforms the other two models across all evaluation metrics. With an accuracy of 88.76%, precision of 87.92%, and recall of 86.48%, the random forest model exhibits superior capability in capturing complex and heterogeneous consumer purchasing behavior patterns. The decision tree model performs better than logistic regression; however, it shows limitations in maintaining prediction stability on the testing dataset.

Beyond numerical superiority, the enhanced performance of the random forest model can be theoretically explained through its ensemble structure, which aggregates multiple decision trees trained on different bootstrap samples. This mechanism reduces variance while maintaining low bias, thereby improving generalization capability. Compared with logistic regression, which assumes linear relationships among variables, random forest effectively captures non-linear interactions between demographic attributes, transaction frequency, and expenditure patterns.

These findings are consistent with previous research indicating that ensemble learning methods outperform single classifiers in retail consumer behavior prediction contexts [15], [21]. For instance, Stylianou and Pantelidou [15] reported improved predictive stability using tree-based ensemble models in supermarket analytics, while Sharipov [21] emphasized the robustness of machine learning ensembles in modeling consumer heterogeneity.

From a business perspective, the superior predictive performance of the random forest model has practical implications for retail information systems. Accurate prediction of purchasing behavior enables more efficient inventory planning by anticipating demand fluctuations, supports targeted promotional strategies through customer behavior forecasting, and enhances personalization mechanisms in digital retail platforms. Therefore, integrating predictive analytics into retail information systems transforms them into proactive decision-support infrastructures rather than passive reporting tools.

Further performance evaluation was conducted using the Receiver Operating Characteristic (ROC) curve to assess the discriminative ability of the random forest model across different threshold levels. The ROC curve is presented in figure 2.

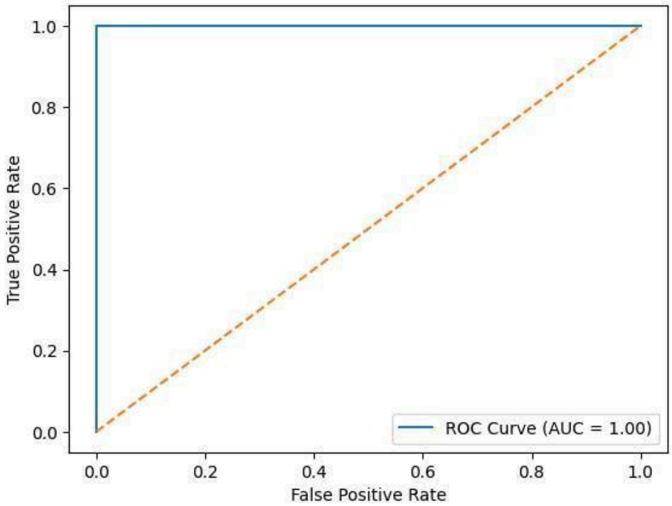

Figure 2. ROC Curve of the Random Forest Model

The ROC curve lies significantly above the diagonal line, indicating strong classification performance. The high Area Under the Curve (AUC) value confirms the model's ability to consistently distinguish between positive and negative consumer purchasing behavior classes.

To gain deeper insight into the performance of the best-performing model, a confusion matrix analysis was conducted for the random forest model, as presented in Table 2.

Table 2. Confusion Matrix of the Random Forest Model

	predicted positive	**predicted negative**
Actual Positive	TP = 432	FN = 68
Actual Negative	FP = 54	TN = 446

The confusion matrix indicates a relatively high number of true positive and true negative predictions compared to misclassification errors. The relatively low values of false positives and false negatives suggest that the random forest model achieves a balanced performance in detecting positive purchasing behavior while minimizing classification errors. This balance is particularly important in retail information systems, where misclassification may directly affect business decision making.

3.3. Discussion

The results of the predictive analytics modeling demonstrate that machine learning–based approaches provide deeper insights into consumer purchasing behavior compared to conventional descriptive analysis. The superior performance of the random forest model, as evidenced by accuracy, precision, recall, and ROC–AUC results, indicates that consumer purchasing behavior exhibits complex and non-linear patterns. This finding is consistent with prior studies reporting that ensemble learning methods are more effective in capturing interactions among multiple behavioral variables.

Compared with logistic regression and decision tree models, the random forest model shows greater prediction stability and lower misclassification rates. This advantage can be attributed to its ensemble structure, which combines multiple decision trees to reduce model bias and variance. In the context of retail information systems, such robustness is critical, as business decisions derived from predictive outcomes require high reliability. Inaccurate predictions may lead to ineffective marketing strategies, suboptimal inventory management, or inappropriate service personalization.

P-ISSN 2085-4315
E-ISSN 2502-8332

Moreover, the findings indicate that integrating predictive analytics into retail information systems extends their role beyond traditional reporting tools toward intelligent decision support systems. This study addresses an existing research gap that often separates technical machine learning modeling from information system implementation. By embedding predictive models within retail information systems, this research offers a more holistic perspective on how transaction data can be leveraged to generate strategic and sustainable business insights.

From a practical perspective, the developed predictive model can assist retail information system managers in anticipating consumer purchasing tendencies, thereby supporting inventory planning, promotional strategy development, and personalized service delivery. From an academic standpoint, this study contributes to the information systems literature by emphasizing the importance of integrating machine learning–based predictive analytics into consumer behavior analysis. Consequently, this research not only improves predictive performance but also provides a comprehensive conceptual framework for the role of predictive analytics in modern retail information systems.

4. CONCLUSION

This study confirms that integrating machine learning–based predictive analytics into retail information systems significantly enhances the understanding and forecasting of consumer purchasing behavior beyond conventional descriptive reporting. Among the evaluated models, random forest demonstrated superior predictive performance, reflecting its ability to capture complex and heterogeneous behavioral patterns in retail transaction data. The findings reinforce the theoretical positioning of predictive analytics as an embedded capability within modern information systems architectures and highlight its strategic role in transforming retail systems into proactive decision-support platforms. Practically, the developed framework supports more accurate inventory management, targeted promotional planning, and personalized customer engagement strategies. However, this study is limited by the use of a secondary Kaggle dataset, which may not fully represent real-time enterprise retail environments. Future research should explore real-world transactional systems, incorporate deep learning approaches, and examine system integration challenges in operational retail infrastructures.

REFERENCES

[1] A. Berisha Qehaja, "Digital Transformation in Retail: Leveraging Data Analytics for Competitive Advantage BT - Data Analytics for Decision Making towards Business Excellence: Through the Lens of Retail and Financial Services," B. Basu, J. K. Jha, I. Mukherjee, and R. N. Sengupta, Eds., Singapore: Springer Nature Singapore, 2026, pp. 15–37. doi: 10.1007/978-981-95-2429-7_2.

[2] D. J. Kim and S. S. Cha, "Digital Transformation in Retail : A Systematic Review on Omnichannel , AI , and Big Data," *Distrib. Sci. Res.*, vol. 6, no. 2024, pp. 111–124, 2025, doi: 10.15722/jds.23.06.202506.111.

[3] M. Kholod, A. Celani, and G. Ciaramella, "The Analysis of Customers' Transactions Based on POS and RFID Data Using Big Data Analytics Tools in the Retail Space of the Future," 2024. doi: 10.3390/app142411567.

[4] T. Stylianou and A. Pantelidou, "Big data and consumer behavior : A macroeconomic perspective through supermarket analytics," *QFE*, vol. 9, no. July, pp. 682–712, 2025, doi: 10.3934/QFE.2025024.

[5] S. Suhairi, N. Nurhazizah, S. Syanda, and R. Nasution, "Transformasi Digital Riset Pemasaran Global dengan Integrasi Teknologi Terkini untuk Menyusun Strategi Responsif terhadap Perubahan Pasar Global," *As-Syirkah Islam. Econ. & Financ. J.*, vol. 3, no. 2 SE-Articles, Jan. 2024, doi: 10.56672/syirkah.v3i2.175.

[6] Loso Judijanto, I. Pratama, I. W. Jata, and E. Utami, "Analisis Bibliometrik tentang Pengaruh Big Data dan Analitik dalam Pengembangan Produk dan Layanan," *J. Multidisiplin West Sci.*, vol. 3, no. 01 SE-Artikel, pp. 88–97, 2024, doi: 10.58812/jmws.v3i01.942.

[7] S. A. Alenazi, "Predictive modelling of omni-channel customer behavior using big data analytics for retail marketing," *Int. J. Innov. Res. Sci. Stud.*, vol. 8, no. 5 SE-Articles, pp. 1350–1359, Aug. 2025, doi: 10.53894/ijirss.v8i5.9134.

[8] J. Tang, "Unlocking Retail Insights: Predictive Modeling and Customer Segmentation Through Data Analytics," 2025. doi: 10.3390/jtaer20020059.

[9] G. Karpushkin, "Predicting Consumer Behavior Based on Big Data of User-Generated Online Content in

Retail Marketing," *Glob. J. Flex. Syst. Manag.*, vol. 25, no. 1, pp. 163–178, 2024, doi: 10.1007/s40171-024-00372-5.

[10] P. Haessner, J. Haessner, and J. Thomas, "Maximizing Retail Potential: The Role of Big Data Analytics," *J. Strateg. Innov. Sustain.*, vol. 19, no. 4 SE-Articles, Jan. 2025, doi: 10.33423/jsis.v19i4.7486.

[11] R. S. Ganesh and R. S. Sagidya, "Business Analytics: Challenges and Innovations for the Modern Enterprise," *Shanlax Int. J. Manag.*, vol. 11, no. 4 SE-Articles, Apr. 2024, doi: 10.34293/management.v11i4.7469.

[12] Robert P Rooderkerk, Nicole DeHoratius, and Andrés Musalem, "The past, present, and future of retail analytics: Insights from a survey of academic research and interviews with practitioners," *Prod. Oper. Manag.*, vol. 31, no. 10, pp. 3727–3748, Oct. 2022, doi: 10.1111/poms.13811.

[13] Abdullah Al Maruf, "A Systematic Review Of Erp-Integrated Decision Support Systems For Financial And Operational Optimization In Global Retails Business," *Am. J. Interdiscip. Stud.*, vol. 6, no. 1 SE-Articles, pp. 236–262, 2025, doi: 10.63125/qgbrmf24.

[14] V. Krishna and P. Mattegunta, "Data-driven retail : The interconnected ecosystem of predictive merchandising analytics," *World J. Adv. Res. Rev.*, vol. 26, no. 01, pp. 4084–4092, 2025, doi: 10.30574/wjarr.2025.26.1.1570.

[15] T. Stylianou and A. Pantelidou, "A machine learning approach to consumer behavior in supermarket analytics," *Decis. Anal. J.*, vol. 16, p. 100600, 2025, doi: 10.1016/j.dajour.2025.100600.

[16] S. Das and J. Nayak, "Customer Segmentation via Data Mining Techniques: State-of-the-Art Review BT - Computational Intelligence in Data Mining," J. Nayak, H. S. Behera, B. Naik, S. Vimal, and D. Pelusi, Eds., Singapore: Springer Nature Singapore, 2022, pp. 489–507.

[17] M. Devega, K. Kusrini, E. Utami, K. A. Yuana, and N. Kapoor, "Transformative Impact of Data Diversity and Machine Learning on Supply Chain Management," in *2025 4th International Conference on Creative Communication and Innovative Technology (ICCIT)*, 2025, pp. 1–6. doi: 10.1109/ICCIT65724.2025.11167560.

[18] C. Magasi and A. M. Nyamwesa, "Reimagining Consumer Analytics: Predictive and Real-Time Insights Through Dynamic Structural Equation Modeling," *Indones. J. Bus. Entrep.*, vol. 11, no. 3 SE-Articles, p. 734, 2025, doi: 10.17358/ijbe.11.3.734.

[19] V. Krishna and P. Mattegunta, "Integrated retail ecosystem : The convergence of predictive analytics and omnichannel strategies in modern merchandising," *World J. Adv. Res. Rev.*, vol. 15, no. 02, pp. 168–176, 2025, doi: 10.30574/wjaets.2025.15.2.0509.

[20] J. Lin, "Application of machine learning in predicting consumer behavior and precision marketing," *PLoS One*, vol. 20, no. 5, p. e0321854, May 2025, doi: 10.1371/journal.pone.0321854.

[21] R. Sharipov, "Machine Learning Systems for Predicting Consumer Behaviour," *Asian J. Res. Comput. Sci.*, vol. 18, no. 6 SE-Review Article, pp. 343–352, Jun. 2025, doi: 10.9734/ajrcos/2025/v18i6704.

[22] K. S. Rekha, T. Amutha, R. Usharani, S. Pushparani, and V. M. Sivagami, "Predictive Analysis of Consumer Behavior Using Supervised Learning Techniques," in *2024 Second International Conference Computational and Characterization Techniques in Engineering & Sciences (IC3TES)*, 2024, pp. 1–6. doi: 10.1109/IC3TES62412.2024.10877475.

[23] N. Alsharman, I. Hababeh, and D. Alshorman, "Leveraging Machine Learning with Feature Selection to Enhance Business Intelligence in Predicting E-Commerce Acquiring Behaviors," *Int. J. Adv. soft Comput. its Appl.*, vol. 17, no. 3, pp. 117–133, 2025, doi: 10.15849/IJASCA.251130.07.

[24] S. S. and S. S., "Intelligent Deep Learning Systems for Predicting Customer Purchase Intentions in Phygital Retail Industry," in *2025 International Conference on Emerging Trends in Industry 4.0 Technologies (ICETI4T)*, 2025, pp. 1–6. doi: 10.1109/ICETI4T63625.2025.11132150.

[25] X. Ma and X. Jiang, "Predicting Cross-border E-commerce Purchase Behavior in Organic Products: A Machine Learning Approach Integrating Cultural Dimensions and Digital Footprints," *Int. J. Comput. Inf. Syst.*, vol. 5, no. 1, 2024, doi: 10.29040/ijcis.v5i1.212.

[26] D. Anwar, M. Faizanuddin, F. Rahman, and R. Dayal, "Analyzing Consumer Behavior in E-Commerce: Insights from Data-Driven Approaches," *Manag.*, vol. 3 SE-Original, p. 127, Mar. 2025, doi: 10.62486/agma2025127.

[27] S. Hossain, H. Hena, and P. Sampa, "Decoding Consumer Habits: Analyzing Retail Patterns Across

Demographics," *Startupreneur Bus. Digit. (SABDA Journal)*, vol. 3, no. 2 SE-Articles, pp. 148–159, Oct. 2024, doi: 10.33050/sabda.v3i2.638.

[28] A. O. Abiodun, O. Adeolu, B. D. Abayomi, and O. A. Julius, "Development of an Artificial Intelligence-Based Model for Patient's Vital Signs Deterioration Prediction," *Int. J. Innov. Sci. Res. Technol.*, vol. 10, no. 9, 2025, doi: 10.38124/ijisrt/25sep1438.

[29] O. E. Aribisala, "Data Quality Assessment and Preprocessing Techniques for Enhancing Machine Learning Model Performance," *Int. J. Innov. Res. Eng. & Multidiscip. Phys. Sci.*, vol. 12, no. 3, pp. 1–11, 2024, doi: 10.37082/ijirmps.v12.i3.232799.

[30] S. Shahidi, A. W. Samadzai, and H. Shahbazi, "Effective Data Preprocessing in Data Science : From Method Selection to Domain-Specific Optimization," *J. Adv. Comput. Knowl. Algorithms*, vol. 2, no. 4, pp. 84–90, 2025, doi: 10.29103/jacka.v2i4.22886.

[31] S. Diwandari, A. E. Permanasari, and I. Hidayah, "Smart Features, Smarter Insights: The Role of Feature Engineering in Customer Purchase Predictions in E-Commerce," in *2025 International Conference on Advanced Machine Learning and Data Science (AMLDS)*, 2025, pp. 142–148. doi: 10.1109/AMLDS63918.2025.11159401.

[32] S. Sowndharya and R. Hariharan, "Advanced Machine Learning Algorithms: A Big Data Approach to Predictive Analytics in Business Marketing," in *2025 4th International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*, 2025, pp. 2083–2088. doi: 10.1109/ICIMIA67127.2025.11200955.

[33] J. I. García-Guerra, H. O. Aguilar-Cajas, H. E. Vergara-Zurita, A. L. Rivera-Abarca, F. Armijos-Arcos, and J. I. López-Pumalema, "Predictive Analytics in Digital Marketing: A Statistical Modeling Approach for Predicting Consumer Behavior," *Data Metadata*, vol. 4 SE-Original, p. 1061, Jun. 2025, doi: 10.56294/dm20251061.

[34] D. Chicco and G. Jurman, "An Invitation to Greater Use of Matthews Correlation Coefficient in Robotics and Artificial Intelligence," *Front. Robot. AI*, vol. Volume 9-2022, 2022, doi: 10.3389/frobt.2022.876814.

[35] J. H. Cabot and E. G. Ross, "Evaluating prediction model performance," *Surgery*, vol. 174, no. 3, pp. 723–726, Sep. 2023, doi: 10.1016/j.surg.2023.05.023.

[36] A. L. D. Araújo *et al.*, "Artificial intelligence in healthcare applications targeting cancer diagnosis—part II: interpreting the model outputs and spotlighting the performance metrics," *Oral Surg. Oral Med. Oral Pathol. Oral Radiol.*, vol. 140, no. 1, pp. 89–99, Jul. 2025, doi: 10.1016/j.oooo.2025.01.002.

[37] J. O. Ogunniyi, J. O. Emuoyibofarhe, J. B. Oladosu, and M. M. Olamoyegun, "Assessment of some selected Machine Learning Performance Metrics in the Prediction of type 2 Diabetes," *FUOYE J. Eng. Technol.*, vol. 9, no. 3 SE-Articles, pp. 452–457, Dec. 2024, doi: 10.4314/fuoyejet.v9i3.12.

[38] I. S. R, M. M, F. C, and J. K. M, "Microplastic predictive modelling with the integration of Artificial Neural Networks and Hidden Markov Models (ANN-HMM)," *J. Environ. Heal. Sci. Eng.*, vol. 22, no. 2, pp. 579–592, 2024, doi: 10.1007/s40201-024-00920-2.