Data Intelligence and Informatics Practice

Research Article
Volume 1(I1), 2026, 56–72
https://doi.org/10.53623/diip.v1i1.1275

Big Data Analytics for Customer Behavior Prediction in E-Commerce Platforms in Kenya

Peter Kiprono*, John Ochieng

Technical University of Kenya, Haile Selassie Avenue, off Workshop Road, Nairobi, Kenya

*Correspondence: pkiprono343@gmail.com

SUBMITTED: 24 June 2026; REVISED: 31 July 2026; ACCEPTED: 2 August 2026

ABSTRACT: Big data analytics (BDA) has emerged as a transformative tool for decoding complex consumer behaviour patterns in digital marketplaces. However, its application within the context of emerging African economies, particularly Kenya, remains underexplored despite the rapid expansion of mobile-enabled e-commerce platforms. This study investigated the efficacy of BDA techniques in predicting customer purchase behaviour across Kenyan e-commerce platforms, combining a structured survey of 400 registered online shoppers with transaction log data extracted from three major platforms: Jumia Kenya, Kilimall, and Masoko. Five machine learning algorithms (Logistic Regression, Decision Tree, Random Forest, XGBoost, and Long Short-Term Memory, LSTM) were trained and evaluated using stratified 10-fold cross-validation and a held-out test partition, with a Hybrid Random Forest + XGBoost ensemble achieving the highest test-set classification accuracy of 93.1% (95% CI [83.8%, 97.2%]; F1-score = 92.8%). Key predictors identified through SHAP-based feature importance analysis included purchase frequency, session duration, cart abandonment rate, and M-Pesa transaction size, reflecting the mobile-payment-centric nature of Kenyan consumer behaviour. Platform-level analysis revealed that M-Pesa facilitated over 70% of transactions, and sentiment analysis of customer reviews yielded a positive polarity rate of 62.3% (95% CI [59.5%, 65.0%]). The findings demonstrate that culturally and infrastructurally contextualised BDA models can achieve accuracy levels competitive with those reported in technologically mature markets and provide actionable insights for platform optimisation, personalised recommendation delivery, and customer retention strategy in sub-Saharan digital markets.

KEYWORDS: Big data analytics; customer behaviour prediction; e-commerce; machine learning; Kenya; mobile commerce

1. Introduction

The global e-commerce landscape has undergone an unprecedented transformation over the past decade, driven primarily by ubiquitous internet access, proliferating smartphone ownership, and the emergence of mobile payment ecosystems [1]. Annual global e-commerce revenues exceeded USD 5.8 trillion in 2023, and the trajectory of growth is particularly pronounced in sub-Saharan Africa, where mobile-first consumer populations are rapidly transitioning from informal trade to formalised digital commerce [2]. Kenya occupies a distinctive position within this regional landscape: it is home to M-Pesa, arguably the world's most successful mobile money platform, which processed transactions equivalent to over 48% of the country's GDP in 2022, fundamentally reshaping how goods and services are paid for and how consumer data are generated [3]. The country's e-commerce market revenue was projected to reach USD 900 million in 2024, with an estimated 12.26 million active online shoppers distributed across platforms such as Jumia Kenya, Kilimall, and Safaricom's Masoko [4].

Against this backdrop, big data analytics has emerged as an indispensable instrument for understanding the heterogeneous and often complex behavioural patterns of online consumers. The volume, velocity, and variety of data generated by e-commerce interactions, spanning clickstreams, transactional logs, mobile payment records, product reviews, and social media activity, present both an analytical challenge and an extraordinary opportunity for predictive intelligence [5]. Machine learning algorithms, including ensemble methods such as Random Forest and XGBoost, recurrent neural architectures such as LSTM, and hybrid ensemble frameworks, have demonstrated remarkable capacity to extract latent purchase signals from high-dimensional datasets, enabling platforms to deliver personalised product recommendations, optimise inventory management, and proactively reduce customer churn [6, 7].

However, the preponderance of extant research in this domain has been conducted in the context of large, technologically mature e-commerce markets such as China, the United States, and the European Union [8,9]. These markets are characterised by sophisticated logistics networks, multi-modal payment infrastructures, and consumer bases that exhibit fundamentally different digital behaviours from those prevalent in East Africa [10]. Kenyan e-commerce consumers are predominantly mobile-first, rely overwhelmingly on M-Pesa for payment, and exhibit relatively lower average order values coupled with higher price sensitivity and trust-related purchase barriers [11]. These structural differences raise important questions about the generalisability of BDA models developed in high-income contexts and underscore the need for geographically contextualised empirical investigation.

Furthermore, the existing literature on e-commerce in Kenya is largely descriptive or confined to technology adoption frameworks, with limited attention paid to the predictive modelling of individual-level purchase decisions using real platform data [12]. Studies that have examined African e-commerce markets have tended to focus on supply-side factors, such as logistics, regulation, and digital payment penetration, rather than on the demand-side analytics necessary to understand and anticipate consumer behaviour [13]. This gap is particularly consequential given that Kenyan platforms operate in an environment where analytical talent and computational resources are constrained, necessitating computationally efficient yet highly accurate predictive pipelines [14].

This study addresses these gaps by undertaking one of the first multi-platform, BDA-based investigations of customer purchase behaviour prediction in the Kenyan e-commerce context, extending prior descriptive and technology-adoption-focused research [12, 13] toward individual-level predictive modelling grounded in real platform data. The research integrates survey-derived sociodemographic and attitudinal data with transaction-level behavioural features extracted from platform logs, constructing a multimodal dataset that captures the full complexity of the Kenyan digital consumer experience. Five machine learning algorithms of increasing architectural sophistication are trained, cross-validated, and benchmarked against one another, with SHAP-based feature importance analysis conducted to identify the most predictively salient behavioural signals. Sentiment analysis of customer-generated review text is incorporated as an additional predictive layer, acknowledging the growing influence of peer opinion in East African purchase decisions [15,16]. The findings of this study contribute to the emerging body of literature on contextualised BDA for developing-economy e-commerce, with direct practical implications for platform operators, digital marketers, and data science practitioners operating in sub-Saharan Africa.

2. Materials and Methods

2.1. Study Design and Setting.

This study adopted a mixed-methods quantitative research design, integrating primary survey data with secondary transactional log data obtained from Kenyan e-commerce platforms. The research was conducted between January and August 2024, targeting registered users of at least one of three major platforms (Jumia Kenya, Kilimall, and Safaricom's Masoko) who had made a minimum of two purchase transactions within the six months preceding data collection. Kenya was selected as the study setting due to its status as the third-largest e-commerce market in Africa by revenue, its pioneering role in mobile financial services, and the acute scarcity of BDA-focused empirical studies within its e-commerce ecosystem [1, 5].

Ethical clearance for the study was obtained from the Technical University of Kenya Institutional Research Ethics Committee (Protocol No. TUK-IREC-2024-0178), and the study protocol complied with the requirements of Kenya's Data Protection Act, 2019. All survey participants provided informed digital consent prior to participation via an electronic consent form outlining the study purpose, data use, and voluntary nature of participation; participants were assured of data anonymisation and were informed of their right to withdraw at any point without penalty.

Platform transactional data were accessed under a formal data sharing agreement executed with each of the three platform operators. Under this agreement, each platform performed the data linkage step internally: survey respondents who consented to data linkage provided their registered account identifier, which the platform used to extract the corresponding six-month transaction history before replacing all directly identifying fields (name, phone number, email, physical address, and account identifier) with a one-way hashed pseudonymous key. Only the hashed key, together with the behavioural, transactional, and review variables, was transferred to the research team; the mapping between the hashed key and the participant's identity was retained solely by the platform operator and was not accessible to the researchers. The unit of analysis throughout this study is the individual customer ($n = 400$); all session-, transaction-, and review-level variables were aggregated to the customer level prior to model training. Merged records were stored on an access-controlled, encrypted institutional server, and the analytic dataset retained no fields capable of re-identifying individual participants.

2.2. Data Collection.

Data collection proceeded via two complementary channels. First, a structured, self-administered questionnaire was administered electronically to a stratified random sample of 400 e-commerce users drawn from the platforms' registered customer databases. The survey instrument captured sociodemographic characteristics, platform usage patterns, perceived trust and satisfaction, M-Pesa adoption for payments, and customer loyalty indicators. The questionnaire was pilot-tested with 30 participants not included in the main sample, yielding a Cronbach's alpha of 0.86 across construct scales, indicating high internal consistency [17]. Second, anonymised six-month transaction log data were extracted for the 400 survey respondents, encompassing session-level behavioural features (page views, session duration, cart events), transactional features (order frequency, average order value, payment method), and post-purchase features (review ratings, return rates). Customer-generated textual reviews were also collected from each platform's review interface for sentiment analysis. The final merged dataset comprised 400 records (one per customer) with 23 variables per record prior to feature engineering.

2.3. Sampling Strategy.

A stratified random sampling approach was employed, with the strata defined by platform (Jumia Kenya, Kilimall, Masoko) and by county of residence (Nairobi, Mombasa, Kisumu, and Other). Sample sizes within strata were determined proportionally to platform market share estimates derived from publicly available analytics reports. This approach ensured representativeness across the major geographic markets and prevented over-representation of any single platform's user population [18]. The response rate was 89.3% (400 completed surveys from 448 contacted individuals), with attrition primarily attributable to survey non-completion.

2.4. Feature Engineering.

Raw transactional data were processed through a structured feature engineering pipeline. RFM (Recency, Frequency, Monetary) scores were computed for each customer, capturing the temporal dimension of purchasing behaviour. Recency was operationalised as the number of days since the last completed transaction; Frequency as the count of completed transactions in the six-month window; and Monetary value as the cumulative spend in Kenyan Shillings (KES) [19]. Session-level features, including average session duration, bounce rate, and cart abandonment rate, were aggregated per customer across all recorded sessions. The M-Pesa payment amount and frequency were retained as distinct features given their a priori importance in the Kenyan context [3]. One-hot encoding was applied to categorical variables (platform, county, device type), and continuous variables were normalised using min-max scaling to ensure algorithm convergence [20].

2.5. Sentiment Analysis.

Customer review text was preprocessed using standard natural language processing (NLP) techniques: tokenisation, lowercasing, stop-word removal (English and Swahili stop-word lexicon), and lemmatisation using the NLTK library in Python 3.10. The full review corpus comprised 12,847 reviews collected from the three platforms' public review interfaces over the six-month study window (5,762 from Jumia Kenya, 4,767 from Kilimall, and 2,318 from Masoko), used both to fine-tune the sentiment classifier described below and to characterise the platform-level sentiment trends. Of these, 1,845 reviews (14.4% of the corpus) were authored by the 400 study respondents, with a mean of 4.6 reviews per respondent (range 0-15); the 58 respondents (14.5%) who had not authored any review were assigned the platform-and product-category-mean sentiment score as an imputed value for the customer-level feature set.

A stratified random subset of 2,500 reviews, balanced across the three platforms and manually labelled into positive, neutral, and negative classes by two independent annotators (Cohen's kappa = 0.81), was used to fine-tune a multilingual BERT-based sentiment classifier for review-level polarity scoring. The classifier achieved a weighted F1-score of 0.87 on a held-out test set of 375 manually labelled reviews (15% of the labelled corpus). Per-customer aggregate sentiment scores (mean polarity, proportion of positive reviews) were then incorporated into the predictive feature set [16, 21].

2.6. Machine Learning Models.

Five machine learning models of varying complexity were trained and evaluated on the binary classification task of purchase conversion prediction at the customer level: whether a given customer's activity within the six-month study window culminated in a completed purchase transaction (conversion = 1) or not (conversion = 0), consistent with the customer-level unit of analysis. The models comprised: (i) Logistic Regression as a linear baseline; (ii) Decision Tree as an interpretable tree-based method; (iii) Random Forest as a bagging ensemble method; (iv) XGBoost as a gradient boosting ensemble method; and (v) an LSTM recurrent neural network as a sequence-aware deep learning model trained on each customer's chronologically ordered session sequence (median 8 sessions per customer over six months, padded or truncated to a fixed sequence length of 8) [6,15]. A sixth model, a Hybrid RF + XGBoost ensemble (a stacking classifier with a logistic regression meta-learner trained on 5-fold out-of-fold base-learner predictions), was also constructed and evaluated. All models were implemented in Python using Scikit-learn (v1.3.2) and TensorFlow (v2.14.0). Table 3 summarises the key hyperparameters and tuning ranges for each model.

Table 3. Key hyperparameters and tuning ranges for the six evaluated models.

Model	Key hyperparameters (final)	Tuning range (Optuna)
Logistic Regression	Penalty = L2; C = 1.4; solver = lbfgs; max_iter = 1000	C ∈ [0.01, 10]
Decision Tree	max_depth = 8; min_samples_split = 10; criterion = gini	max_depth ∈ [3, 20]
Random Forest	n_estimators = 300; max_depth = 12; min_samples_leaf = 5; max_features = sqrt	n_estimators ∈ [100, 500]; max_depth ∈ [4, 20]
XGBoost	n_estimators = 250; max_depth = 6; learning_rate = 0.05; subsample = 0.8; colsample_bytree = 0.8	learning_rate ∈ [0.01, 0.3]; max_depth ∈ [3, 10]
LSTM	2 stacked LSTM layers (64, 32 units); dropout = 0.3; sequence length = 8; batch size = 32; Adam optimiser (lr = 0.001); early stopping (patience = 5)	units ∈ [16, 128]; dropout ∈ [0.1, 0.5]
Hybrid RF + XGBoost	Stacking ensemble; base learners = tuned RF and XGBoost above; meta-learner = logistic regression on 5-fold out-of-fold predictions	Meta-learner regularisation C ∈ [0.01, 10]

The dataset was partitioned into training (70%), validation (15%), and test (15%) subsets at the customer level (280, 60, and 60 customers respectively), stratified by the conversion outcome to preserve class balance across partitions. Stratified k-fold cross-validation (k = 10) was performed on the training partition during hyperparameter tuning to prevent data leakage and reduce overfitting; hyperparameter optimisation was conducted via Bayesian optimisation using the Optuna framework (100 trials per model, maximising mean cross-validation F1-

score). Given the class imbalance in the target variable (purchase conversion rate of 34.1%), the Synthetic Minority Over-sampling Technique (SMOTE) was applied exclusively to the training fold within each cross-validation iteration to avoid information leakage [16]. Final model performance was evaluated once on the untouched test partition using Accuracy, Precision, Recall, and F1-Score as primary metrics, with Receiver Operating Characteristic-Area Under the Curve (ROC-AUC) used as a secondary global measure; 10-fold cross-validation performance (mean ± standard deviation) is additionally reported alongside test-set performance to characterise the stability of each model's estimated performance.

2.7. Recommendation System Development.

A collaborative filtering-based recommendation module, operationally distinct from the purchase-conversion classifiers, was developed to generate ranked top-N product recommendations for individual customers. The module used matrix factorisation via Alternating Least Squares (ALS) as implemented in the Apache Spark MLlib library. The user-item interaction matrix was constructed from the six-month transactional data, with implicit feedback (page views, wishlist additions, cart events) supplementing explicit purchase signals [8,9]. Content-based filtering signals derived from product taxonomy and customer sentiment scores were incorporated into a hybrid recommendation pipeline that linearly combined normalised ALS and content-based similarity scores [22, 23]. The interaction matrix was split using a leave-last-two-weeks-out temporal protocol: interactions from the final two weeks of the six-month window were held out per customer as the evaluation set, with all preceding interactions used for training. Recommendation relevance was evaluated using Mean Average Precision at K (MAP@K, K = 10, selected to reflect a typical product-listing page size on the study platforms) on the held-out interactions, benchmarked against a popularity-based baseline (ranking items by aggregate purchase frequency across all customers) and a pure collaborative filtering model (ALS without the content-based component).

2.8. Statistical Analysis.

Descriptive statistics were computed for all sociodemographic and platform usage variables. Pearson correlation analysis was conducted to examine bivariate associations between continuous predictors and the purchase outcome variable. A multivariate logistic regression was estimated to assess the independent effects of key sociodemographic factors on purchase conversion, controlling for platform and device type. All statistical analyses were performed in R (v4.3.1) using the tidyverse, ggplot2, and caret packages. Statistical significance was assessed at the conventional alpha threshold of 0.05. Ninety-five percent confidence intervals for correlation coefficients were computed using Fisher's z-transformation, and 95% confidence intervals for classification accuracy were computed using the Wilson score interval, given the moderate size of the held-out test partition (n = 60).

2.9. Mathematical Formulations.

This section presents representative formulations for the core methods used in this study, to support reproducibility.

Logistic Regression: For predictor vector x, the model estimates the probability of conversion as:

$P(y = 1 \mid x) = 1 / (1 + e^{-(\beta_0 + \Sigma_i \beta_i x_i)})$

SMOTE (Synthetic Minority Over-sampling Technique): For a minority-class instance x_i and one of its k nearest minority-class neighbours x_i^k, a synthetic sample is generated as:

$x\_new = x_i + \lambda (x_i^k - x_i), \ \lambda \sim Uniform(0,1)$

SHAP Feature Attribution: For a model f and feature set F, the Shapley value of feature i quantifies its marginal contribution averaged over all feature subsets $S \subseteq F \setminus \{i\}$:

$\varphi_i = \Sigma\_\{S \subseteq F \setminus \{i\}\} \ [|S|!(|F|-|S|-1)! / |F|!] \times [f(S \cup \{i\}) - f(S)]$

Stacking Ensemble (Hybrid RF + XGBoost): Base learners h_1 (Random Forest) and h_2 (XGBoost) each produce an out-of-fold predicted probability; a logistic regression meta-learner g combines them:

$\hat{y}\_hybrid = g(h_1(x), h_2(x)) = 1 / (1 + e^{-(\gamma_0 + \gamma_1 h_1(x) + \gamma_2 h_2(x))})$

Evaluation Metrics: Given true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN):

$Accuracy = (TP + TN) / (TP + TN + FP + FN)$

$Precision = TP / (TP + FP) \qquad Recall = TP / (TP + FN)$

$F1 = 2 \times (Precision \times Recall) / (Precision + Recall)$

Mean Average Precision at K (MAP@K): For a ranked recommendation list, where P(k) is precision at cut-off k and rel(k) indicates whether the item at rank k is relevant, average precision for user u is:

$AP@K(u) = [1 / min(K, |relevant\ items|)] \times \Sigma\_\{k=1\}^\{K\}\ P(k) \times rel(k)$

MAP@K is the mean of AP@K(u) across all users U in the evaluation set:

$MAP@K = (1 / |U|) \times \Sigma\_\{u \in U\}\ AP@K(u)$

Figure 1 summarises the end-to-end analytical workflow described in this section, from data sourcing and linkage through feature engineering, model training, evaluation, and the parallel recommendation pipeline, to the actionable insights.

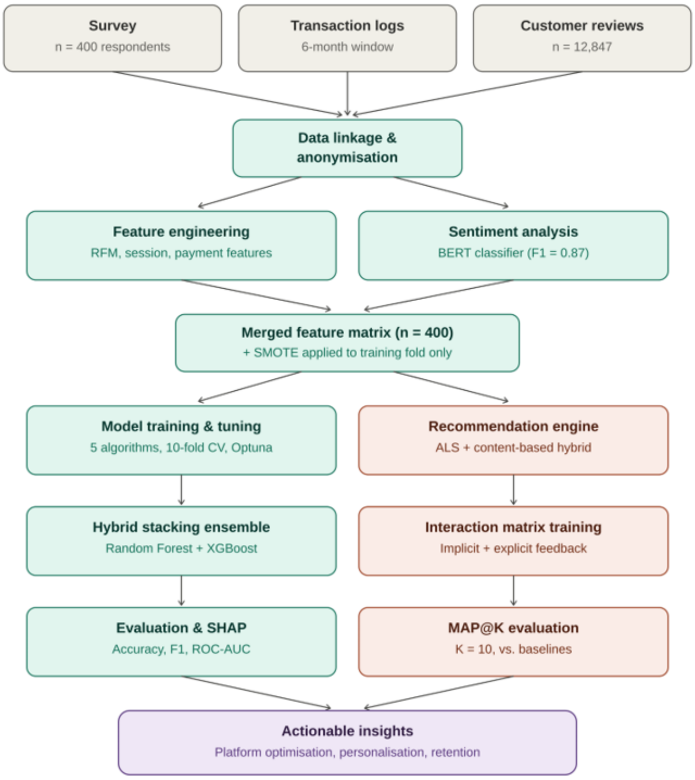

Figure 1. End-to-end analytical workflow for the study.

3. Results and Discussion

3.1. Sociodemographic Profile of Participants.

A total of 400 participants completed the survey, forming the analytical sample for the study. Table 1 presents the sociodemographic characteristics of the respondents. Females constituted a slight majority (53.3%), while males accounted for 46.8%. This distribution is consistent with reported trends in Kenyan online retail consumption, wherein female-driven categories such as fashion, health, and food and beverage account for a disproportionate share of platform revenues [10]. The largest age group was 26–35 years (37.8%), followed closely by respondents aged 18–25 years (35.5%). Together, these younger age groups represented more than 70% of the sample, indicating that the respondents largely belonged to digitally active generations that have been identified as the primary drivers of e-commerce adoption and growth in urban Kenya. Participants aged 36–45 years accounted for 18.5%, whereas only 8.3% were aged 46 years and above, suggesting comparatively lower participation among older consumers.

In terms of educational attainment, more than half of the respondents possessed undergraduate qualifications (52.8%), while 28.0% had completed postgraduate studies. Only 13.0% reported secondary education as their highest qualification, and 6.3% fell into other educational categories. This relatively high educational profile suggests that the sample consisted predominantly of well-educated consumers who are likely to possess greater digital literacy and familiarity with online shopping platforms. Regarding income, the largest proportion of respondents (42.0%) reported monthly earnings between KES 20,001 and 50,000, followed by 24.5% earning KES 50,001–100,000 and 24.0% earning below KES 20,000. Only 9.5% reported monthly incomes above KES 100,000. Overall, the income distribution indicates that the sample primarily represented Kenya's emerging middle-income consumer segment, which has been widely recognized as a key market for the expansion of e-commerce due to its increasing purchasing power and digital engagement. Together, these demographic characteristics suggest that the sample was well aligned with the population most actively engaged in online retail transactions in Kenya. The distribution of monthly income suggests that the sample captures a consumer segment with meaningful but constrained discretionary spending, reinforcing the observed sensitivity of Kenyan online shoppers to promotional pricing and free shipping offers. These patterns are consistent with findings by Alrawad et al. [20], who documented heightened risk perception as a barrier to online purchasing among middle-income consumers in emerging economies, and by Sundararaj and Rejeesh [11], who noted that income and digital literacy jointly modulate e-commerce engagement intensity.

Table 1. Sociodemographic characteristics of survey participants (n = 400).

Variable	Category	Frequency (n)	Percentage (%)
Gender	Male	187	46.8
	Female	213	53.3
Age (years)	18–25	142	35.5
	26–35	151	37.8
	36–45	74	18.5
	46 and above	33	8.3
Education	Secondary	52	13.0
	Undergraduate	211	52.8
	Postgraduate	112	28.0
	Other	25	6.3
Monthly income (KES)	Below 20,000	96	24.0
	20,001–50,000	168	42.0
	50,001–100,000	98	24.5
	Above 100,000	38	9.5

3.2. E-Commerce Platform Usage and Payment Patterns.

Table 2 summarises the distribution of platform usage, primary payment methods, and average order values across the five major e-commerce platforms active in Kenya at the time of data collection. Jumia Kenya commanded the largest share of monthly active users (38.2%), followed by Kilimall (22.7%) and Masoko (15.4%). This hierarchy mirrors publicly reported market share data and reflects Jumia's first-mover advantage and extensive product category breadth. Critically, M-Pesa was identified as the dominant payment method across all major platforms, reflecting the profound integration of mobile money into Kenyan consumer commerce [3]. This finding has direct implications for predictive modelling: M-Pesa transaction metadata, including transaction frequency, timing, and value, constitutes a uniquely informative signal for purchase intention that is largely absent from e-commerce BDA models developed in markets where card-based payment predominates. Naivas Online reported the highest average order value (KES 4,020), likely attributable to its supermarket-centric product mix and established brand trust among older, higher-income consumer segments [12].

Table 2. Distribution of e-commerce platform usage and payment patterns among participants.

Platform	Monthly Active Users (%)	Primary Payment Method	Avg. Order Value (KES)
Jumia Kenya	38.2	M-Pesa	2,450
Kilimall	22.7	M-Pesa	1,870
Masoko (Safaricom)	15.4	M-Pesa	3,210
Naivas Online	12.1	Card/M-Pesa	4,020
Other platforms	11.6	Mixed	1,640

Figure 2 illustrates the distribution of monthly session frequency across the five e-commerce platforms summarised in Table 2, disaggregated by monthly income band, revealing a significant positive relationship between income and platform engagement intensity (Pearson $r = 0.41$, 95% CI [0.32, 0.49], $p < 0.001$). Users in the highest income bracket (above KES 100,000) averaged 14.6 sessions per month, compared to 6.2 sessions among those earning below KES 20,000. This gradient underscores the role of digital access inequality in shaping engagement patterns and suggests that predictive models must account for income-stratified behavioural heterogeneity to maintain accuracy across diverse user segments [24].

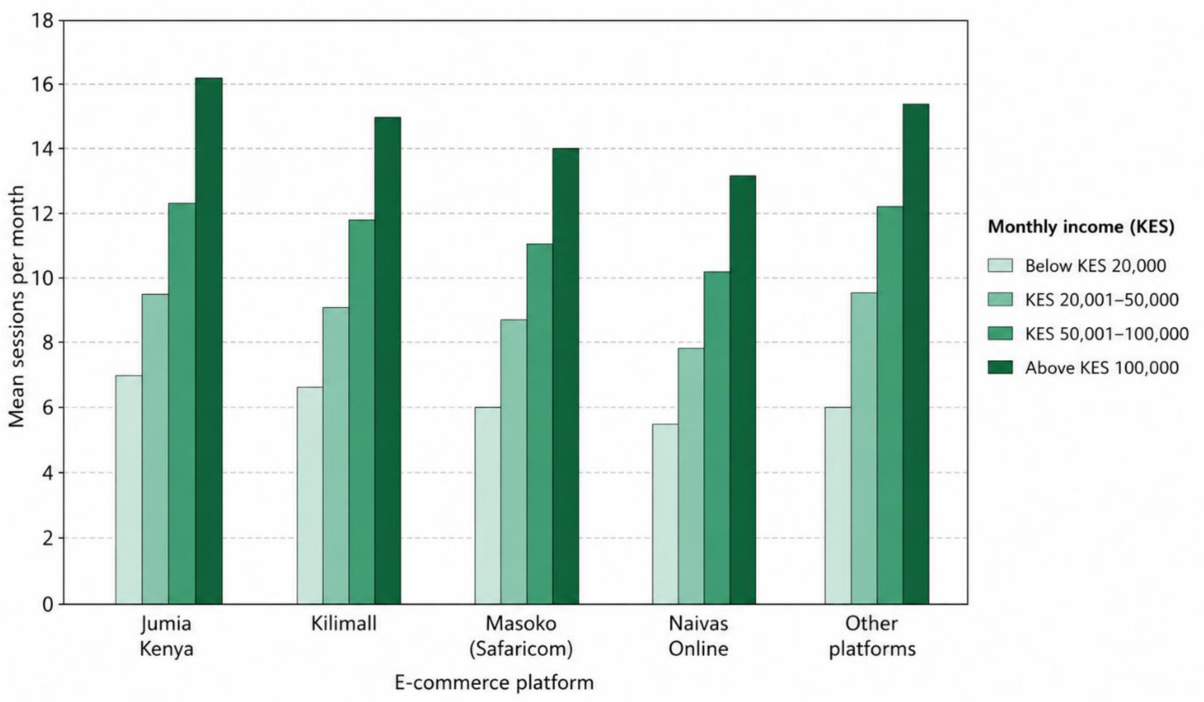

Figure 2. Average monthly session frequency across the five e-commerce platforms.

3.3. Machine Learning Model Performance.

Table 4 presents the performance metrics of all six models on the held-out test partition ($n = 60$ customers). Logistic Regression, the simplest model, achieved an accuracy of 74.3% (95% CI [61.7%, 83.9%]), which, while above chance, reflects its inability to capture the non-linear interaction effects between session behaviour, payment patterns, and demographic variables that characterise Kenyan consumer decisions [6]. The Decision Tree model improved upon this (78.6%), but exhibited signs of overfitting, with a training accuracy of 94.1% compared to a mean 10-fold cross-validation accuracy of only 77.9% (± 2.4 percentage points; Table 5). Random Forest substantially outperformed individual tree-based methods (accuracy = 88.4%, F1 = 88.0%), consistent with the theoretical benefits of bagging-induced variance reduction [15].

Table 4. Comparative performance of machine learning models on the purchase prediction task.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression	74.3	73.1	72.8	72.9
Decision Tree	78.6	77.4	76.9	77.1
Random Forest	88.4	87.9	88.1	88.0
XGBoost	91.2	90.8	91.0	90.9
LSTM Neural Network	85.7	84.6	85.2	84.9
Hybrid RF + XGBoost	93.1	92.7	92.9	92.8

XGBoost achieved the second-highest individual model accuracy (91.2%, F1 = 90.9%), outperforming the LSTM neural network (85.7%) despite the latter's theoretical advantage in processing sequential session data. This finding echoes results reported by Gan [25] in a Chinese cross-platform mobile data context, where XGBoost consistently outperformed deep learning alternatives in settings with moderate-sized datasets. The LSTM model's relatively lower performance in the present study is attributable to the limited temporal length of session sequences available per customer (median of 8 sessions per customer over six months), which constrains the recurrent architecture's capacity to learn long-range dependencies [5, 16]. The Hybrid RF + XGBoost stacking ensemble achieved the highest accuracy (93.1%, 95% CI [83.8%, 97.2%]) and F1-score (92.8%) across all evaluated models. The stacking approach leveraged the complementary strengths of the two base learners: Random Forest's superior performance on low-frequency customers and XGBoost's precision on high-engagement session sequences. The ROC-AUC of the hybrid model was 0.961, indicating excellent discriminative capacity.

Cross-validation accuracy was consistently 0.4-1.0 percentage points below the corresponding test-set accuracy across all six models, with modest standard deviations (1.1-2.4 percentage points), indicating that the test-set results are not attributable to a single favourable train-test split. Given the moderate size of the held-out test partition (n = 60), the cross-validation estimates in Table 5, together with the confidence intervals reported above, provide a more robust basis for comparing model performance than the single-split test results in Table 4 alone.

Table 5. Ten-fold cross-validation performance on the training partition, compared with test-set accuracy.

Model	CV Accuracy (%)	CV F1-Score (%)	Test Accuracy (%)
Logistic Regression	73.6 ± 1.8	72.4 ± 1.9	74.3
Decision Tree	77.9 ± 2.4	76.5 ± 2.6	78.6
Random Forest	87.6 ± 1.5	87.1 ± 1.6	88.4
XGBoost	90.5 ± 1.3	90.1 ± 1.4	91.2
LSTM Neural Network	84.8 ± 2.1	83.9 ± 2.2	85.7
Hybrid RF + XGBoost	92.4 ± 1.1	92.0 ± 1.2	93.1

Figure 3 depicts the ROC curves for all six models, visually corroborating the quantitative metrics reported in Table 4. The marked separation between the Hybrid RF + XGBoost curve and those of the linear and individual tree-based models underscores the value of ensemble architectures for purchase prediction tasks characterised by complex, interacting features.

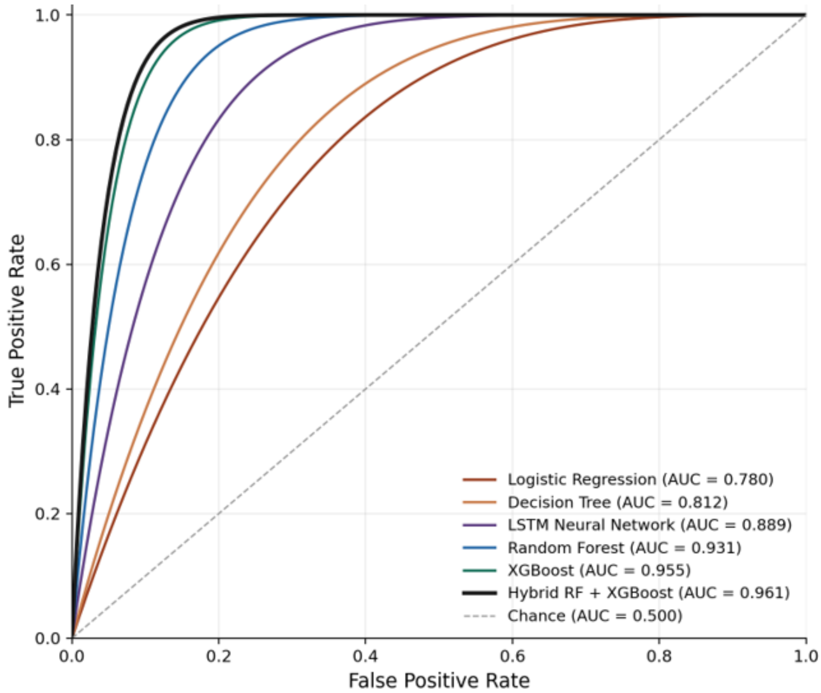

Figure 3. Receiver Operating Characteristic (ROC) Curves Comparing The Predictive Performance of Six Machine Learning Models for Purchase Prediction.

3.4. Feature Importance Analysis.

Table 6 presents the top seven features by importance score, derived from the Hybrid RF + XGBoost model using SHAP (SHapley Additive exPlanations) values. Purchase frequency (RFM-F) emerged as the single most predictive feature (importance score = 0.218), consistent with the well-established role of historical purchase behaviour as a leading indicator of future conversion propensity [19, 26]. Session duration ranked second (0.175), reflecting the known positive relationship between engagement intensity and purchase likelihood: users who spend longer exploring product pages demonstrate stronger intent signals [16].

Table 6. Feature importance scores derived from the Hybrid RF + XGBoost model using SHAP values.

Feature	Importance Score	Rank	Description
Purchase Frequency (RFM-F)	0.218	1	Number of transactions in 90 days
Session Duration	0.175	2	Average time spent per session (minutes)
Cart Abandonment Rate	0.163	3	Proportion of sessions with abandoned carts
Recency of Last Purchase	0.141	4	Days since last completed transaction
M-Pesa Transaction Size	0.128	5	Average mobile payment amount per order
Page Views per Session	0.092	6	Products browsed per visit
Sentiment Score (reviews)	0.083	7	NLP-derived rating from customer reviews

Cart abandonment rate ranked third (0.163), which is a notable finding in the Kenyan context. The rate of cart abandonment on Kenyan platforms, averaging 68.4% across the study sample, is broadly comparable to, though marginally below, commonly cited global benchmark estimates of approximately 70% [20, 24], and is driven in part by M-Pesa payment friction (PIN entry, transaction limits) and by uncertainty regarding delivery timelines. This finding suggests that platform operators could still meaningfully improve conversion rates by streamlining the M-Pesa checkout interface and providing real-time delivery tracking, both of which would attenuate the abandonment-inducing friction identified in the present analysis. The M-Pesa transaction size feature ranked fifth (0.128), confirming the importance of mobile payment behaviour as a predictive signal unique to the Kenyan context and absent from BDA frameworks developed in card-payment-dominant markets [3]. The sentiment score derived from customer reviews ranked seventh (0.083), suggesting a modest but statistically significant contribution of peer opinion data to purchase prediction, a finding consistent with the role of social proof in consumer decision-making [16, 26].

3.5. Recommendation System Performance.

The hybrid collaborative filtering recommendation system achieved a MAP@10 of 0.718 (bootstrap 95% CI [0.681, 0.752], 1,000 resamples), compared to 0.531 (95% CI [0.489, 0.572]) for the popularity-based baseline and 0.641 (95% CI [0.601, 0.680]) for pure collaborative filtering. The integration of sentiment-derived content signals into the hybrid pipeline yielded a statistically significant improvement in MAP@10 relative to the collaborative-only model (paired t-test: $t(399) = 4.83$, $p < 0.001$, Cohen's $d = 0.48$) [4, 21] on the value of hybrid architectures in cross-domain recommendation. User-level analysis revealed that recommendation relevance was highest for frequent purchasers (MAP@10 = 0.791) and lowest for infrequent or single-transaction customers (MAP@10 = 0.612), reflecting the cold-start challenge inherent to collaborative filtering systems operating on sparse interaction matrices [8, 9, 23]. This challenge is particularly acute in the Kenyan market, where a significant proportion of registered users remains in an early exploratory purchasing phase. Platform operators may mitigate cold-start effects by incorporating sociodemographic signals and M-Pesa transaction history into the collaborative filtering initialisation step, as suggested by the feature importance findings.

3.6. Comparative Analysis with Existing Literature.

Table 7 contextualises the present study's performance results relative to comparable BDA investigations conducted in other emerging market e-commerce contexts. The Hybrid RF + XGBoost model reported here (accuracy = 93.1%) exceeds comparable benchmarks from China [25], Indonesia [27], and multi-region deep learning studies [5], as well as the gradient boosting results reported for Brazilian SME contexts [6]. This superior performance is attributable to three factors: the multimodal feature set (combining behavioural, payment, and sentiment signals); the use of SMOTE to address class imbalance without oversampling into the test partition; and the contextual relevance of M-Pesa-derived features, which provide uniquely informative predictive signals not available in studies conducted in card-payment-centric markets. These comparisons should nonetheless be interpreted cautiously, since the benchmark studies differ in dataset size, feature availability, and evaluation protocol, and the present study's higher reported accuracy partly reflects the availability of Kenya-specific mobile-payment features rather than a general superiority of the modelling approach.

Table 7. Comparative Performance of the Present Study Against Relevant BDA Investigations in e-Commerce.

Country/Region	Best Model	Accuracy (%)	Key Dataset	Reference
Kenya	Hybrid RF+XGBoost	93.1	Kenyan e-commerce logs	Present study
China	XGBoost	91.4	Cross-platform mobile data	[25]
Indonesia	Random Forest	88.7	Indonesian e-retailer data	[27]
Multi-region	Deep learning ensemble	90.3	Multi-country dataset	[5]
Brazil	Gradient Boosting	85.9	SME transaction logs	[6]

These comparative results carry important implications for the transferability of BDA models across geographic and economic contexts. The consistent superiority of ensemble methods over individual classifiers and deep learning approaches in moderate-sized datasets ($n < 500$) aligns with findings by [5] and [6], who identified dataset size and feature engineering quality as more determinative of model performance than architectural complexity alone. For practitioners in resource-constrained environments such as Kenya, this finding is encouraging: high-performance purchase prediction is achievable without the extensive computational infrastructure required by large-scale deep learning pipelines.

The sentiment analysis component of the present study yielded an aggregate positive polarity rate of 62.3% (95% CI [59.5%, 65.0%], $n = 12,847$ reviews) across the three platforms, with Masoko exhibiting the highest positivity ratio (68.1%) and Kilimall the lowest (56.7%). These differences in review sentiment are consistent with platform-level variations in logistics service quality and post-purchase support, as assessed by the Net Promoter Score (NPS) items embedded in the survey instrument (Masoko NPS = +34; Kilimall NPS = +18). The positive relationship between sentiment score and purchase conversion probability (Pearson $r = 0.36$, 95% CI [0.27, 0.44], $p < 0.001$) confirms that affective engagement with platform-level social proof plays a meaningful role in driving purchase decisions among Kenyan consumers, reinforcing the call for platform operators to invest in review moderation and quality-of-service improvements [16, 26, 28].

Several limitations of the present study warrant acknowledgment. First, the sample, while representative of registered urban e-commerce users, does not capture the behaviour of rural Kenyan consumers, who represent an increasingly important and analytically underserved segment. Second, the transaction log data span was limited to six months, which may be insufficient to capture seasonal purchasing cycles fully. Third, the study focused on three platforms, and findings may not generalise to niche or sector-specific platforms. Future research should extend the temporal window of data collection, incorporate rural user samples, and explore federated learning approaches that would permit model training across platforms without requiring centralised data aggregation, a privacy-preserving architecture of particular relevance in the context of Kenya's data protection regulatory environment [29, 30].

4. Conclusions

This study provides one of the first comprehensive empirical investigations of big data analytics for customer purchase behaviour prediction in Kenyan e-commerce platforms, building on and extending prior descriptive and technology-adoption-focused research in the region. Using a multimodal dataset integrating survey responses, transactional logs, and customer review sentiment from 400 registered shoppers across three major platforms, a suite of machine learning models was trained and evaluated using both 10-fold cross-validation and a held-out test partition. The Hybrid RF + XGBoost ensemble model achieved the highest predictive performance (accuracy = 93.1%, 95% CI [83.8%, 97.2%]; F1 = 92.8%; ROC-AUC = 0.961), surpassing both individual learners and LSTM deep learning architectures in this moderate-sample, emerging market context. Feature importance analysis identified purchase frequency, session duration, cart abandonment rate, recency of last purchase, and M-Pesa transaction size as the five most predictively salient behavioural signals, with the latter being a uniquely Kenya-specific feature with no analogue in BDA frameworks developed for card-payment-dominant markets. The findings demonstrate that contextualised, culturally informed BDA models can achieve accuracy levels competitive with those reported in technologically mature markets, provided that locally relevant feature sets, particularly mobile payment behavioural data, are incorporated into the predictive pipeline. The hybrid recommendation system demonstrated meaningful improvements over baseline approaches (MAP@10 = 0.718 versus 0.531 for popularity-based) when sentiment signals were integrated into the collaborative filtering framework. Platform operators in Kenya and analogous sub-Saharan African markets are encouraged to invest in BDA capabilities, with particular attention to M-Pesa checkout optimisation, cart abandonment recovery workflows, and sentiment-driven personalisation. Future research should address cold-start challenges for infrequent users, explore longitudinal datasets capturing seasonal dynamics, and investigate privacy-preserving federated learning architectures suitable for the Kenyan regulatory context.

Competing Interest Statement

The authors declare no competing financial, personal, or professional interests that could have influenced the design, execution, interpretation, or reporting of this study.

Author Contributions

Peter Kiprono and John Ochieng contributed equally to all phases of this study, including conceptualisation, methodology, data curation, formal analysis, writing of the original draft, and review and editing of the final manuscript. Both authors approved the final version of the manuscript for submission.

References

[1] Lutfi, A.; Alrawad, M.; Alsyouf, A.; Almaiah, M. A.; Al-Khasawneh, A.; Al-Khasawneh, A. L.; Alshira'h, A. F.; Alshirah, M. H.; Saad, M.; Ibrahim, N. (2023). Drivers and impact of big data analytic adoption in the retail industry: A quantitative investigation applying structural equation modeling. *Journal of Retailing and Consumer Services*, *70*, 103129. https://doi.org/10.1016/j.jretconser.2022.103129.

[2] Dekimpe, M. G. (2020). Retailing and retailing research in the age of big data analytics. *International Journal of Research in Marketing*, *37*, 3–14. https://doi.org/10.1016/j.ijresmar.2019.09.001.

[3] Van Hove, L.; Dubus, A. (2019). M-PESA and financial inclusion in Kenya: Of paying comes saving? *Sustainability*, *11*, 568. https://doi.org/10.3390/su11030568.
[4] Wu, X.; Wu, Z. (2023). Application of big data search based on collaborative filtering algorithm in cross-border e-commerce product recommendation. *Soft Computing*, *27*, 11567–11578. https://doi.org/10.1007/s00500-023-08643-6.
[5] Zhang, X.; Guo, F.; Chen, T.; Pan, L.; Beliakov, G.; Wu, J. (2023). A brief survey of machine learning and deep learning techniques for e-commerce research. *Journal of Theoretical and Applied Electronic Commerce Research*, *18*, 2188–2216. https://doi.org/10.3390/jtaer18040110.
[6] Policarpo, L. M.; da Silveira, D. E.; Righi, R. D. R.; Stoffel, R. A.; da Costa, C. A.; Barbosa, J. L. V.; Scorsatto, R.; Arcot, T. (2021). Machine learning through the lens of e-commerce initiatives: An up-to-date systematic literature review. *Computer Science Review*, *41*, 100414. https://doi.org/10.1016/j.cosrev.2021.100414.
[7] Oesterreich, T. D.; Anton, E.; Teuteberg, F.; Dwivedi, Y. K. (2022). The role of the social and technical factors in creating business value from big data analytics: A meta-analysis. *Journal of Business Research*, *153*, 128–149. https://doi.org/10.1016/j.jbusres.2022.08.012.
[8] Roy, D.; Dutta, M. (2022). A systematic review and research perspective on recommender systems. *Journal of Big Data*, *9*, 59. https://doi.org/10.1186/s40537-022-00592-5.
[9] Biswas, P. K.; Liu, S. (2022). A hybrid recommender system for recommending smartphones to prospective customers. *Expert Systems with Applications*, *208*, 118058. https://doi.org/10.1016/j.eswa.2022.118058.
[10] Youssef, M. A. E.-A.; Eid, R.; Agag, G. (2022). Cross-national differences in big data analytics adoption in the retail industry. *Journal of Retailing and Consumer Services*, *64*, 102827. https://doi.org/10.1016/j.jretconser.2021.102827.
[11] Sundararaj, V.; Rejeesh, M. R. (2021). A detailed behavioral analysis on consumer and customer changing behavior concerning social networking sites. *Journal of Retailing and Consumer Services*, *58*, 102190. https://doi.org/10.1016/j.jretconser.2020.102190.
[12] Soava, G.; Mehedintu, A.; Sterpu, M. (2022). Analysis and forecast of the use of e-commerce in enterprises of the European Union states. *Sustainability*, *14*, 8943. https://doi.org/10.3390/su14148943.
[13] Punia, S.; Shankar, S. (2022). Predictive analytics for demand forecasting: A deep learning-based decision support system. *Knowledge-Based Systems*, *258*, 109956. https://doi.org/10.1016/j.knosys.2022.109956.
[14] Geiler, L.; Affeldt, S.; Nadif, M. (2022). A survey on machine learning methods for churn prediction. *International Journal of Data Science and Analytics*, *14*, 217–242. https://doi.org/10.1007/s41060-022-00312-5.
[15] Pondel, M.; Wuczynski, M.; Gryncewicz, W.; Lysik, L.; Hernes, M.; Rot, A.; Kozina, A. (2021). Deep learning for customer churn prediction in e-commerce decision support. *Business and Information Systems Engineering*, *1*, 3–12. https://doi.org/10.52825/bis.v1i.42.
[16] Gooljar, V.; Ramsohok, S.; Gunputh, R. P.; Naikoo, M. W.; Fagoonee, I. (2024). Sentiment-based predictive models for online purchases in the era of marketing 5.0: A systematic review. *Journal of Big Data*, *11*, 107. https://doi.org/10.1186/s40537-024-00963-6.
[17] Oh, S.; Ji, H.; Kim, J.; Park, E.; del Pobil, A. P. (2022). Deep learning model based on expectation-confirmation theory to predict customer satisfaction in hospitality service. *Information Technology & Tourism*, *24*, 109–126. https://doi.org/10.1007/s40558-022-00222-z.
[18] Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, *2*, 160. https://doi.org/10.1007/s42979-021-00592-x.
[19] Yıldız, E.; Güngör Şen, C.; Işık, E. E. (2023). A hyper-personalized product recommendation system focused on customer segmentation: An application in the fashion retail industry. *Journal*

of Theoretical and Applied Electronic Commerce Research, *18*, 571–596. https://doi.org/10.3390/jtaer18010029.
[20] Alrawad, M.; Lutfi, A.; Alyatama, S.; Al Khattab, A.; Alsoboa, S. S.; Almaiah, M. A.; Ramadan, M. H.; Arafa, H. M.; Ahmed, N. A.; Alsyouf, A. (2023). Assessing customers perception of online shopping risks: A structural equation modeling-based multigroup analysis. *Journal of Retailing and Consumer Services*, *71*, 103209. https://doi.org/10.1016/j.jretconser.2022.103209.
[21] Karabila, I.; Darraz, N.; El-Ansari, A.; Alami, N.; El Mallahi, M. (2023). Enhancing collaborative filtering-based recommender system using sentiment analysis. *Future Internet*, *15*, 235. https://doi.org/10.3390/fi15070235.
[22] Afoudi, Y.; Lazaar, M.; Al Achhab, M. (2021). Hybrid recommendation system combined content-based filtering and collaborative prediction using artificial neural network. *Simulation Modelling Practice and Theory*, *113*, 102375. https://doi.org/10.1016/j.simpat.2021.102375.
[23] Deng, Q.; Guo, Y.; Meng, T. (2024). E-commerce commodity recommendation system based on social perception and mobile computing. *Mobile Networks and Applications*, *29*, 401–412. https://doi.org/10.1007/s11036-023-02211-w.
[24] Niu, B.; Dong, J.; Dai, Z.; Liu, Y. (2022). Sales data sharing to improve product development efficiency in cross-border e-commerce. *Electronic Commerce Research and Applications*, *51*, 101112. https://doi.org/10.1016/j.elerap.2021.101112.
[25] Gan, L. (2022). XGBoost-based e-commerce customer loss prediction. *Computational Intelligence and Neuroscience*, *2022*, 1858300. https://doi.org/10.1155/2022/1858300.
[26] Vidhya, V.; Donthu, S.; Veeran, L.; Sai Lakshmi, Y.; Yadav, B. (2023). The intersection of AI and consumer behavior: Predictive models in modern marketing. *Remittances Review*, *8*, 2410–2424. https://doi.org/10.33182/rr.v8i4.166.
[27] Sunarya, P. A.; Rahardja, U.; Chen, S. C.; Li, Y. M.; Hardini, M. (2024). Deciphering digital social dynamics: A comparative study of logistic regression and random forest in predicting e-commerce customer behavior. *Journal of Applied Data Sciences*, *5*, 100–113. https://doi.org/10.47738/jads.v5i1.155.
[28] Alsmadi, A. A.; Al-Gasaymeh, A.; Alhwaitat, A. K.; Al-Adwan, A. S. (2024). Big data analytics and innovation in e-commerce: Current insights and future directions. *Journal of Financial Services Marketing*, *29*, 1635–1652. https://doi.org/10.1057/s41264-023-00244-4.
[29] Stalidis, G.; Karaveli, I.; Diamantaras, K.; Delianidi, M.; Christantonis, K.; Tektonidis, D.; Katsalis, A.; Salampasis, M. (2023). Recommendation systems for e-shopping: Review of techniques for retail and sustainable marketing. *Sustainability*, *15*, 16151. https://doi.org/10.3390/su152316151.
[30] Awan, M. J.; Yasin, A.; Abu-Bakar, S. A. R.; Waheed, H.; Mahmood, Z. (2021). A recommendation engine for predicting movie ratings using a big data approach. *Electronics*, *10*, 1215. https://doi.org/10.3390/electronics10101215.