Vol 8, Issue 6, 2025 **Impact Factor: 8.535** DOI: https://doi.org/10.31426/ijamsr.2025.8.6.8416

IJAMSR

International Journal of Advanced Multidisciplinary Scientific Research (IJAMSR) ISSN:2581-4281

Advanced Machine Learning and Data Mining Techniques for Analysing Consumer Behavior Patterns in Large-Scale Transactional, Textual, and Rating Datasets

Rupam Kumari [1]

[1] Research Scholar, Department of Computer Science, Capital University, Koderma, Jharkhand.

Prof. (Dr.) Anand Kumar [2]

[2] Supervisor, Department of Computer Science, Capital University, Koderma, Jharkhand.

Dr. Supriya Shree [3]

[3] Co-Supervisor, Assistant Professor, Department of Computer Science,
St. Xavier's College of Management & Technology, Patna.

ABSTRACT

This study explores advanced machine learning and data mining techniques to analyse consumer behaviour patterns using large-scale transactional, textual, and rating datasets. Through integrating psychological, emotional, and behavioural dimensions, the research demonstrates how consumer decisions are shaped by cultural, social, and affective factors expressed through digital footprints. Using the Online Retail dataset and a 1M-record electronics rating dataset, the study applies clustering, PCA, text mining, collaborative filtering, and multiple classification models. Results show strong long-tail patterns, distinct product and customer segments, and superior predictive performance from ensemble models particularly XGBoost, Random Forest, and a Voting Classifier. Emotion-aware recommender architectures further highlight the potential of incorporating sentiment and social-context signals. The findings emphasize that hybrid, data-driven approaches can significantly enhance segmentation, forecasting, and personalized recommendations, offering valuable insights for intelligent marketing and consumer analytics.

Keywords: *Consumer Behaviour, Machine Learning, Data Mining, Recommender Systems.*

1. Introduction

Advanced data mining techniques play a transformative role in understanding and predicting consumer behaviour by uncovering deep insights from large, complex datasets. One of the most powerful methods is Association Rule Mining (ARM), which identifies meaningful relationships among products frequently purchased together. Algorithms such as Apriori, FP-Growth, and Eclat help retailers design effective cross-selling strategies, product bundles, and optimized store layouts.

Modern ARM models also integrate machine learning to capture seasonal and non-linear purchasing associations. Another important technique is advanced clustering, which uses algorithms like K-means++, DBSCAN, OPTICS, and Self-Organizing Maps (SOM) to create refined customer segments. These methods detect hidden patterns, outliers, and non-linear boundaries within consumer data, allowing businesses to design highly personalized marketing campaigns, loyalty programs, and targeted promotions. When combined with predictive modelling, clustering can also identify customers transitioning toward higher-value segments. Sequential Pattern Mining (SPM) further enhances behavioural understanding by analysing the order in which consumers make purchases over time. Algorithms like PrefixSpan and SPADE identify recurring buying sequences, enabling businesses to predict future purchases, design lifecycle marketing strategies, and manage inventory more effectively. In addition, text mining and opinion mining apply Natural Language Processing techniques such as TF-IDF, LDA, and Word2Vec to extract insights from reviews, feedback, and social media content. This helps organizations assess brand sentiment, customer satisfaction, and emerging product preferences. Finally, combining predictive and prescriptive analytics enables businesses not only to forecast consumer actions such as churn or purchase probability—but also to determine the best strategic response [1-8].

2. Related Reviews

Garg et al. (2025, June) had investigated hybrid machine learning models for forecasting customers' purchasing behaviour using e-commerce data. They had combined RF, XGBoost and SVM with ANN to capture complex patterns in browsing and transaction records. The data were reportedly cleaned through missing-value treatment, normalization, and categorical encoding. Model performance had been evaluated using accuracy, precision, recall, F1-score, and AUC-ROC. The hybrid XGBoost–ANN model was said to have achieved the best performance, with 97% accuracy at epoch 30, followed by SVM–ANN (96%) and RF–ANN (95%). The study had highlighted the superiority of hybrid approaches for accurate, scalable e-commerce behaviour prediction.

Koyluoglu and Esme (2025) had explored how ML applications were reshaping marketing by predicting digital consumer actions and delivering the right content at the right time. Their study had focused on modelling the relationship between consumer buying behaviour (CBB) and consumption metaphor (CM). Two scenarios were created: one correlating CBB with CM for validation, and another using ML to predict CBB and assess CM's influence. KNN reportedly achieved 91.02% accuracy for predicting consumers and 90.98% for non-intenders regarding tattoos. When CM was included, prediction accuracies around 78–79% had confirmed that psychological metaphors could be effectively linked to purchasing patterns through ML.

Navarro (2024) had undertaken an exploratory study on how ML-driven predictive analytics was transforming business understanding of customer behaviour. The research had reviewed algorithms, data types and practical implementations used to derive behavioural insights. Findings suggested that firms applying ML-based analytics could refine marketing strategies, optimize customer service, and improve overall consumer experience. The study had also addressed ethical challenges such as privacy, transparency and responsibility in data-driven targeting. It was concluded that, with expanding data and technological advances, predictive analytics held substantial potential for strengthening consumer engagement and granting firms a sustained competitive advantage in increasingly digital markets.

Panduro-Ramirez (2024, May) had aimed to develop an ML-based approach for analysing customer behaviour on e-commerce platforms. Recognizing the rapid growth of online shopping, the study had used clustering, classification, and predictive modelling on large-scale transactional, interactional, and demographic data. This framework reportedly identified distinct customer segments, revealed purchasing patterns, and anticipated future actions. The research suggested that such ML-driven analytics could significantly enhance marketing effectiveness, customer satisfaction, and strategic decision-making. Overall, the study had contributed a structured methodology showing how e-commerce firms might leverage data-driven models to understand evolving consumer behaviour and design more targeted, responsive marketing interventions.

Necula (2023) had examined how time spent reading product information influenced online consumer behaviour. Focusing on navigation patterns in e-commerce environments, the study had used ML techniques and clickstream analysis to uncover hidden structures in browsing data. Customer clusters were identified, and non-linear relationships among variables such as reading duration, bounce rate, exit rate and customer type were explored. Findings suggested that longer engagement with product information, combined with specific behavioural indicators, significantly affected purchase likelihood. The study had offered methodological contributions for modelling complex digital behaviours and provided practical implications for improving website design, content presentation, and targeted marketing strategies.

Sarabhai et al. (2023, November) had investigated the intersection of AI, ML, and behavioural economics to predict consumer buying behaviour, with a focus on alcoholic beverage consumption. Using 384 respondents aged 20–40 in Tashkent, the study had captured views from students, professionals, and entrepreneurs. It had analysed how bounded rationality, cognitive factors (stress, concentration), social and emotional pressures, and physical challenges shaped purchasing decisions. Results indicated that these behavioural and contextual variables significantly influenced buying and consumption patterns. The work had extended consumer analytics by integrating behavioural economics constructs with AI/ML, offering richer explanations of complex purchase decisions.

Hicham and Karim (2022) had considered how rising competition and similar service offerings pushed firms to focus on service quality and personalization. Their study had proposed using ML to better understand clients, design individualized services and improve marketing effectiveness. They implemented six algorithms—Random Forest, Gradient Boosting, Logistic Regression, LightGBM, XGBoost and Decision Tree—to predict consumer behaviour and support targeted campaigns. Comparative results reportedly showed Gradient Boosting as the best-performing and most efficient model. The authors had concluded that ML-assisted analytics could enhance customer acquisition, strengthen relationships, and support long-term loyalty in service-dominated markets.

Mitchell (2022) had discussed how ML had transformed consumer behaviour analysis by enabling accurate forecasting of purchasing tendencies and highly personalized marketing actions. Drawing on big data, deep learning and NLP, the study had described how firms could extract insights from interactions, social media, and purchase records. Key applications reportedly included recommendation systems, churn prediction, sentiment analysis and predictive marketing analytics. The paper had also highlighted challenges such as data privacy, algorithmic bias, and interpretability of complex models. Emerging directions were said to include more transparent, ethical, and explainable AI frameworks to balance commercial value with consumer protection.

Zou (2021) had explored the complex drivers behind consumer resale behaviour and proposed a measurement model using ML and BP neural networks. Through incorporating a contraction–expansion factor and differential evolution operator, the model balanced global and local search and enhanced population diversity. Survey data were collected, trained, and tested within this framework, and real versus predicted values were compared through graphical visualizations. Results indicated that the model effectively captured resale behaviour patterns and provided reliable predictions. The study had been positioned as a theoretical reference for future research on modelling and understanding consumers' resale decisions in online environments.

Anshu et al. (2021, October) had focused on predicting consumer behaviour in online marketing during the COVID-19 period. Using an Amazon dataset from Kaggle containing reviews, ratings, and product attributes, they applied several ML models. Their proposed Random Forest classifier had achieved outstanding performance, with an accuracy of about 98.73%. Comparative analysis with other algorithms confirmed its superiority. The study had demonstrated that robust feature engineering and ensemble models could deliver highly accurate forecasts of consumer intentions, thereby supporting more effective targeting, recommendation, and strategic planning in e-commerce contexts under rapidly changing market conditions.

Raza et al. (2020) had developed a low-cost, non-intrusive ML-based method to quantify consumer-behaviour-driven energy wastage (CBB-EW) in HVAC operations. Using temperature and humidity sensors, they modelled environmental and HVAC-related heat flows to infer ON/OFF status and energy use. CBB-EW was divided into non-occupancy-based and occupancy-based components, identified through data fusion with motion and contextual information and analysed via the PMV comfort model. An experimental office case study showed that users often wasted more than 50% energy through unnecessary operation and sub-optimal settings. The work had highlighted the need for personalized feedback and behavioural interventions for energy conservation.

Juárez-Varón et al. (2020) had examined which aspects of package design most strongly influenced consumer attention when purchasing educational toys. Using neuromarketing experiments, they developed an ML-based methodology to predict which packaging areas were first viewed and which were ignored. Data from eye-tracking and related measures were analysed with advanced models to segment packaging zones by consumer type and social context. Findings indicated that graphic details were the most influential elements guiding visual attention. The proposed approach had demonstrated how neuromarketing and ML could optimize communication design, improving packaging effectiveness and aligning visual cues with consumer preferences.

Choudhury and Nur (2019) had argued that understanding purchasing behaviour was vital for revenue growth and competitiveness. They proposed an ML-based framework to identify potential customers for a retail superstore, replacing traditional policy-based statistical approaches. Using engineered features from historical purchase records, they built classification models to distinguish high-potential customers. After testing multiple algorithms, the study reported an impressive prediction accuracy of 99.4%. The findings had demonstrated that ML techniques could uncover subtle patterns missed by conventional methods and support more precise targeting, resource allocation, and strategic planning in retail environments.

Bayoude et al. (2018) had reviewed how digital marketing was evolving amid new tools, shifting consumer habits and exploding data volumes. They observed that marketers often struggled to extract clear insights from massive, fragmented datasets. The study highlighted that around 84% of marketing organizations were adopting or expanding ML use in 2018. Their work had mapped the state of the art in ML techniques for digital marketing and illustrated how these tools could be scaled to analyse large datasets. They concluded that integrating ML enabled deeper understanding of target audiences and more optimized, evidence-based marketing interactions.

Singh and Tucker (2017) had investigated how online product reviews could be exploited using ML to aid both consumers and designers. They noted that ratings and textual feedback on e-commerce sites offered rich information but were difficult to process manually and sometimes misinterpreted. Their framework classified reviews into direct product characteristics (form, function, behaviour) and indirect ones (service, other), and compared algorithms for this task. High accuracies (around 79–82%) were reported across multiple validation settings. The analysis showed that form had the strongest correlation with star ratings, suggesting that physical design heavily influenced overall customer evaluations.

Cominola et al. (2016) had addressed rising household water demand due to urban expansion and stressed the need for tailored demand-management strategies. They proposed a data-driven methodology integrating clustering and ML to model and classify household water consumption profiles. Using real consumption datasets, the study identified heterogeneous user groups and characterized distinct usage patterns. The results demonstrated that such profiling could reveal underlying behavioural structures and priority segments for conservation. The authors concluded that their approach supported personalized demand-management policies, targeted educational campaigns and collaborative actions between water utilities and consumers for more sustainable urban water use.

Crawford et al. (2015) had surveyed ML techniques for detecting “opinion spam” in online reviews, recognizing that fake feedback could distort consumer decisions and market outcomes. They described how NLP-derived textual features and reviewer metadata could be fed into supervised classifiers to identify deceptive content. The review compared prominent algorithms and highlighted heavy reliance on labelled datasets, which were often scarce. The authors also noted that Big Data tools had been underutilized despite rapidly growing review volumes. Their work had provided a comprehensive comparative overview and outlined future research directions for more scalable, robust spam-detection frameworks.

Karg (2014) had examined service selection in future markets where many functionally equivalent options existed. The study emphasized that non-functional attributes such as response time, price and availability were critical differentiators, yet SLAs were often static and unreliable due to provider embellishment. Within the SOC project, the researchers developed a service broker framework that monitored real service behaviour and used ML to predict non-functional characteristics. A foreground/background architecture separated real-time requests from learning processes. Simulated real-world evaluation indicated 70% accuracy for selecting the best-fit service and 94% for the top two, supporting scalable, context-aware service selection.

Motycka et al. (2013) had investigated the classification of marketing research data using ML techniques. They applied three algorithms ZeroR, J48 Decision Tree and PART to a survey dataset on consumer behaviour in the Czech food market. Each method was evaluated for accuracy, interpretability, and suitability for marketing analytics. The comparative analysis identified which model produced the best balance between predictive performance and explanatory clarity. Their findings had illustrated how ML could enhance the analysis of marketing survey data, improve segmentation and target, and support more informed managerial decision-making within competitive consumer markets.

Calvert and Brammer (2012) had discussed how ML applied to fMRI data could revolutionize marketing communication pretesting. By analysing brain responses to brands, products and prior campaigns, pattern-recognition algorithms reportedly identified neural signatures associated with success or failure. New campaigns or products could then be evaluated by comparing their neural patterns to those of proven campaigns, reducing reliance on subjective verbal feedback. The authors argued that this approach offered faster, more objective, and investment-justified predictions of consumer acceptance. They suggested that such neuroimaging-based ML tools might significantly reshape how marketers design and validate persuasive communications.

3. Findings of Studies

Author(s) & Year	Objective / Focus	Data & Methods Used	Key Findings / Outcomes
Garg et al. (2025)	To forecast customer purchasing behaviour using hybrid ML models	E-commerce transaction and browsing data; RF, XGBoost, SVM combined with ANN; evaluated using accuracy, precision, recall, F1 and AUC-ROC	Hybrid XGBoost–ANN achieved the best performance (97% accuracy), followed by SVM–ANN (96%) and RF–ANN (95%), confirming the superiority of hybrid models
Koyluoglu & Esme (2025)	To predict digital consumer actions and assess the influence of consumption metaphors	Consumer buying behaviour (CBB) and consumption metaphor (CM) data; KNN-based ML models	KNN achieved over 91% accuracy for predicting purchase intention; inclusion of CM confirmed psychological metaphors could be linked to purchasing patterns
Navarro (2024)	To explore the role of ML-driven predictive analytics in understanding customer behaviour	Conceptual and empirical review of ML algorithms, data types and business applications	ML analytics improved marketing strategies, customer service and experience, while raising ethical concerns regarding privacy and transparency

Panduro-Ramirez (2024)	To develop an ML-based framework for analysing e-commerce customer behaviour	Large-scale transactional, interactional and demographic data; clustering, classification and predictive modelling	The framework identified distinct customer segments and purchasing patterns, enhancing marketing effectiveness and strategic decision-making
Necula (2023)	To analyse the effect of reading time on online purchasing behaviour	Clickstream and navigation data; ML-based clustering and non-linear modelling	Longer engagement with product information significantly increased purchase likelihood, offering insights for website design optimization
Sarabhai et al. (2023)	To integrate behavioural economics with AI/ML for predicting buying behaviour	Survey data (384 respondents); AI/ML models incorporating cognitive and social variables	Behavioural and contextual factors significantly influenced purchase decisions, extending consumer analytics beyond transactional data
Hicham & Karim (2022)	To improve service personalization and marketing effectiveness using ML	Random Forest, Gradient Boosting, Logistic Regression, LightGBM, XGBoost and Decision Tree	Gradient Boosting emerged as the most efficient and accurate model for predicting consumer behaviour
Mitchell (2022)	To examine ML's role in consumer behaviour forecasting and personalization	Big data analytics, deep learning, and NLP techniques	ML enabled recommendation systems, churn prediction, and sentiment analysis, though concerns about bias and interpretability remained
Zou (2021)	To model consumer resale behaviour using ML and neural networks	Survey data; BP neural networks with evolutionary optimization	The proposed model accurately captured resale behaviour and improved prediction reliability
Anshu et al. (2021)	To predict online consumer behaviour during COVID-19	Amazon Kaggle dataset; Random Forest and comparative ML models	Random Forest achieved very high accuracy (98.73%), demonstrating the strength of ensemble models
Raza et al. (2020)	To quantify consumer-behaviour-driven energy wastage	Sensor data from HVAC systems; ML and PMV comfort model	Over 50% energy wastage was attributed to consumer behaviour, highlighting the need for behavioural interventions
Juárez-Varón et al. (2020)	To predict consumer visual attention in package design	Eye-tracking data; neuromarketing experiments with ML	Graphic elements most strongly influenced consumer attention, enabling optimized packaging design
Choudhury & Nur (2019)	To identify potential retail customers using ML	Retail transaction data; classification models	ML models achieved up to 99.4% accuracy, outperforming traditional statistical approaches
Bayoude et al. (2018)	To review ML adoption in digital marketing	Industry reports and literature review	ML adoption was rapidly increasing, enabling scalable and evidence-based marketing analytics
Singh & Tucker (2017)	To classify online product reviews using ML	Textual review and rating data; NLP-based classifiers	Classification accuracies of 79–82% were reported, with product form strongly influencing ratings

Cominola et al. (2016)	To model household water-use behaviour using ML	Real consumption datasets; clustering and classification	ML identified heterogeneous consumption profiles, supporting personalized demand-management strategies
Crawford et al. (2015)	To detect opinion spam in online reviews	Textual and reviewer metadata; supervised ML classifiers	ML was effective for spam detection, though limited labelled data posed challenges
Karg (2014)	To predict service quality attributes in service-oriented markets	Monitoring data; ML-based service broker framework	Prediction accuracy reached 70% for best service and 94% for top two services
Motycka et al. (2013)	To classify marketing survey data using ML	Consumer survey data; ZeroR, J48 and PART algorithms	ML improved segmentation accuracy while balancing interpretability
Calvert & Brammer (2012)	To apply ML on fMRI data for marketing communication testing	Neuroimaging data; pattern-recognition algorithms	ML-based neuro-analysis provided objective prediction of campaign success

Source: Secondary Data

4. Research Methodology

The research on analysing consumer behaviour using machine learning focuses on understanding how large-scale digital interactions generated across e-commerce sites, social media platforms, and mobile applications shape purchasing decisions. The study aims to examine online consumer behaviour (Amazon, Flipkart), analyse it using multiple ML algorithms, and propose a predictive framework [9-15].

4.1 Research Design

The research design involves defining clear questions, selecting multi-source datasets (surveys, reviews, clickstream, social media), preprocessing data, applying algorithms such as decision trees, logistic regression, clustering, and validating models using accuracy, precision, recall, F1-score, and AUC. Recommendation systems play a key role by combining explicit (ratings) and implicit (browsing patterns) data, while collaborative filtering—both memory-based and model-based—predicts user preferences through user–item matrices.

4.2 Data Set Characteristics

The present study utilizes a large-scale business dataset characterized as multivariate, sequential, and time-series in nature, making it suitable for advanced analytical and predictive modelling. The dataset comprises 541,909 instances, providing a rich empirical foundation for machine learning applications. It contains 8 attributes, represented through both integer and real-valued variables, which capture diverse aspects of consumer or transactional behaviour. The dataset has been widely employed for classification and clustering tasks, enabling researchers to build models that can categorize patterns, segment customers, and uncover hidden structures in the data. Its size, structure, and attribute composition make it highly appropriate for evaluating and comparing various machine learning algorithms in the context of consumer behaviour analysis [15-20].

4.3 Tools

Python was used as the primary analytical tool for this study due to its flexibility and strong ecosystem for data science. Libraries such as Pandas and NumPy supported data preprocessing, while Scikit-learn and XGBoost enabled clustering, classification, model evaluation, and visualization, ensuring reproducible, efficient, and scalable machine learning analysis.

4.4 Data Source

The dataset used in this study was obtained from **Kaggle and UCI**, specifically the *Online Retail (UCI) Transnational Dataset*. It contains 541,909 transaction records from a UK-based non-store gift retailer for the period 2010–2011 and is widely used for research on customer behaviour, retail analytics, and machine learning–based predictive modelling.

5. Data Analysis

This study presents a comprehensive machine learning–driven analysis of the Online Retail (UCI) transnational dataset to uncover meaningful patterns in customer purchasing behaviour. Using exploratory data analysis, text mining, clustering, dimensionality reduction, and classification techniques, the research examines product structures, customer segments, and predictive performance across multiple models. Comparative evaluation demonstrates the effectiveness of ensemble and tree-based approaches, particularly Gradient Boosted Decision Trees and Voting Classifiers, in capturing complex non-linear relationships. Temporal validation and recommender-system experiments further strengthen the practical relevance of the findings, highlighting the value of data-driven analytics for customer personalization, predictive reliability, and informed decision-making in real-world e-commerce environments.

This study uses the Online Retail (UCI) transnational dataset containing 541,909 transaction records from a UK-based non-store gift retailer (2010–2011). Each record includes InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, and Country. Most transactions come from the UK, revealing strong geographic concentration. Order-value distribution shows that mid-range baskets (£200–£500) dominate, while very low and very high orders are relatively rare. Text mining of product descriptions reveals a power-law distribution of keywords, with frequent terms like “set,” “heart,” “vintage,” “gift,” “bag,” and “design,” reflecting the firm’s home-decor and gifting focus. K-Means clustering on product–word matrices identify five product clusters ranging from mass-market decorative items and seasonal Christmas goods to utility/craft items and premium jewelry/art pieces. PCA on product features shows high dimensionality, while PCA on customer features indicates that 5–6 components explain most behavioural variance. Customer-level clustering produces 10 distinct segments, including high-value, frequent buyers and several smaller niche or low-activity groups.

Table 1: Accuracy Finding of GBDT Model

S. No.	Algorithm	Accuracy / Precision (%)
1	Linear SVM (LinearSVC)	83.10
2	Logistic Regression	89.75
3	K-Nearest Neighbours (KNN)	86.01
4	Decision Tree	90.03
5	Random Forest	92.80
6	AdaBoost	56.51
7	XGBoost (GBDT)	94.04
8	Voting Classifier (RF + XGB + KNN + LR)	93.35

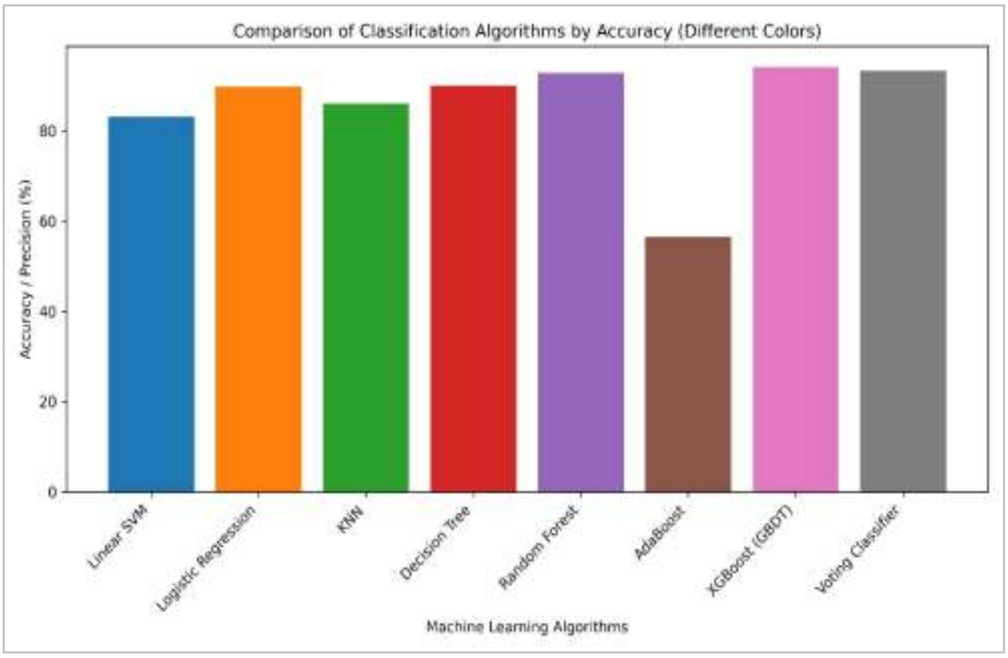

Fig 1: Accuracy Comparison of Machine Learning Classification Algorithms on the Online Retail (UCI) Dataset

It is observed in fig1, that performance among different models vary when applying machine learning algorithms. Considering all types of classifiers, GBDT model (XGBoost) produces the best result with an accuracy of 94.04%, indicating its high ability to capture complex non-linear patterns by gradient boosting. The Voting Classifier (93.35%) based on the combination of Random Forest, XGBoost, KNN and Logistic Regression yields very similar results suggesting that hybrid ensamewhatable results are also present. The methods significantly contribute for predictive dependability.

Several classifiers (Linear SVM, Logistic Regression, KNN, Decision Tree, Random Forest, AdaBoost, XGBoost/GBDT, and a Voting Ensemble) are compared. XGBoost achieves the best accuracy (94.04%), followed by the Voting Classifier (93.35%) and Random Forest (92.80%), while AdaBoost performs weakest (56.51%). Learning curves and confusion matrices show that ensemble models generalize best and separate major clusters clearly, though small or overlapping segments remain challenging. Temporal validation on future data (from 01-10-2011 onward) shows the Voting

Classifier retains 89.22% accuracy, confirming strong robustness over time. Collaborative-filtering and recommender-system experiments further demonstrate how interaction data can be transformed into practical, personalized product suggestions.

The data analysis of the Online Retail (UCI) transnational dataset demonstrates the strong potential of machine learning techniques for extracting actionable business insights from large-scale transactional data. Exploratory analysis revealed clear structural patterns, including geographic concentration in the UK, dominance of mid-range order values, and a product portfolio strongly oriented toward gifting and home décor. Text mining and clustering uncovered meaningful product groupings, while PCA highlighted the inherent high dimensionality of product attributes and more compact customer behavioral patterns. Customer segmentation identified ten distinct behavioral clusters, ranging from high-value frequent buyers to low-activity niche segments, enabling targeted marketing and personalization strategies. Comparative evaluation of multiple classifiers showed that ensemble and tree-based models significantly outperform simpler methods. Among all models, the Gradient Boosted Decision Tree (XGBoost) achieved the highest accuracy of **94.04%**, confirming its superior ability to model complex non-linear relationships. The Voting Classifier closely followed with **93.35%**, indicating that hybrid ensemble approaches offer robust and reliable performance. Temporal validation further confirmed model stability, with strong predictive accuracy on future unseen data. Overall, the study validates the effectiveness of advanced ensemble learning and recommender-system techniques in retail analytics, supporting data-driven decision-making, customer personalization, and scalable predictive modeling in real-world e-commerce environments.

6. Conclusion and Future Scope

The present study on “Advanced machine learning methods and data mining techniques for analysing consumer behaviour patterns” concluded that consumer decisions are shaped not only by price and utility but also by emotions, social influence, culture, and psychological factors, all of which are indirectly reflected in digital footprints such as transactions, ratings, and social-media interactions. Using the Online Retail and electronics rating datasets, the work showed strong long-tail, sparse, and biased structures, revealed clear product and customer segments through text mining, PCA, and clustering, and demonstrated that ensemble models—especially XGBoost, Random Forest, and a Voting Classifier—offer the highest accuracy, robustness, and temporal generalization for customer segmentation and prediction, outperforming simpler linear, KNN, and weak-learner methods. Collaborative filtering and recommender-system experiments further confirmed that raw interaction data can be transformed into personalized, behaviour-aware recommendations, while an emotion-based RS architecture highlighted the potential of integrating affective and social signals into future systems. Looking ahead, the future scope includes deeper integration of multimodal and emotional signals, cross-channel and cross-domain behaviour modeling, advanced representation learning (autoencoders, self-supervision, GNNs), and dynamic sequential and causal models for time-aware decisions. It also calls for hybrid and context-aware recommenders, scalable real-time deployment, and a strong focus on fairness, ethics, explainability, and cross-cultural validation, ensuring that powerful consumer analytics not only improve business performance but also remain transparent, responsible, and human-centered in diverse market environments.

References

1. Anshu, K., Singh, S. K., & Kumari, R. (2021, October). A Machine Learning Model for Effective Consumer Behaviour Prediction. In *2021 5th International Conference on Information Systems and Computer Networks (ISCON)* (pp. 1-5). IEEE.
2. Bayoude, K., Ouassit, Y., Ardchir, S., & Azouazi, M. (2018). How machine learning potentials are transforming the practice of digital marketing: State of the art. *Periodicals of Engineering and Natural Sciences*, *6*(2), 373-379.
3. Calvert, G. A., & Brammer, M. J. (2012). Predicting consumer behavior: using novel mind-reading approaches. *IEEE pulse*, *3*(3), 38-41.
4. Choudhury, A. M., & Nur, K. (2019, January). A machine learning approach to identify potential customer based on purchase behavior. In *2019 international conference on robotics, electrical and signal processing techniques (ICREST)* (pp. 242-247). IEEE.
5. Cominola, A., Giuliani, M., Castelletti, A. F., Spang, E., & Lund, J. (2016). Unveiling Residential Water Consumers' Behaviour and Profiles Through Machine Learning Techniques. In *Proceedings of the World Environmental & Water Resources Congress 2015*. American Society of Civil Engineering (ASCE).
6. Crawford, M., Khoshgoftaar, T. M., Prusa, J. D., Richter, A. N., & Al Najada, H. (2015). Survey of review spam detection using machine learning techniques. *Journal of Big Data*, *2*(1), 23.
7. Garg, K. K., Keshari, N., Biswas, S., Majumder, J., Gangopadhyay, S., & Singh, J. (2025, June). Predicting Consumer Buying Behavior Using Hybrid Machine Learning Models: A Multi-dimensional Analysis of Online Retail Data. In *International Conference on Data Analytics & Management* (pp. 139-150). Cham: Springer Nature Switzerland.
8. Hicham, N., & Karim, S. (2022, October). Machine learning applications for consumer behavior prediction. In *The Proceedings of the International Conference on Smart City Applications* (pp. 666-675). Cham: Springer International Publishing.
9. Juárez-Varón, D., Tur-Viñes, V., Rabasa-Dolado, A., & Polotskaya, K. (2020). An adaptive machine learning methodology applied to neuromarketing analysis: prediction of consumer behaviour regarding the key elements of the packaging design of an educational toy. *Social Sciences*, *9*(9), 162.
10. Karg, P. (2014). Evaluation and Implementation of Machine Learning Methods for an Optimized Web Service Selection in a Future Service Market.
11. Koyluoglu, A. S., & Esme, E. (2025). Predicting the relationship between consumer buying behavior (CBB) and consumption metaphor (CM) through machine learning (ML). *Management & Marketing*, *20*(1).
12. Mitchell, L. (2022). The Use of Machine Learning in Analyzing Consumer Behavior. *American Journal of Data Science and Analysis*, *3*(4), 11-15.
13. MOTYCKA, A., STASTNY, J., & TURCINEK, P. (2013) Marketing Research Data Classification by Means of Machine Learning Methods.

14. Navarro, L. F. M. (2024). Machine Learning and Customer Behavior Insights: Exploring the Depth of Predictive Analytics in Enhancing Consumer Interaction and Engagement. *Journal of Empirical Social Science Studies*, *8*(2), 51-62.
15. Necula, S. C. (2023). Exploring the impact of time spent reading product information on e-commerce websites: A machine learning approach to analyze consumer behavior. *Behavioral Sciences*, *13*(6), 439.
16. Panduro-Ramirez, J. (2024, May). Machine Learning-Based customer behavior analysis for E-commerce platforms. In *2024 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI)* (pp. 1-5). IEEE.
17. Raza, R., Hassan, N. U., & Yuen, C. (2020). Determination of consumer behavior based energy wastage using IoT and machine learning. *Energy and Buildings*, *220*, 110060.
18. Sarabhai, S., Chakraborty, M., Batra, M., Kler, R., Banerjee, S., & Mishra, S. (2023, November). Using AI and Machine Learning to Predict Consumer Buying Behavior: Insights from Behavioral Economics in Case of Alcoholic Beverages. In *2023 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS)* (pp. 980-986). IEEE.
19. Singh, A., & Tucker, C. S. (2017). A machine learning approach to product review disambiguation based on function, form and behavior classification. *Decision Support Systems*, *97*, 81-91.
20. Zou, X. (2021). Analysis of consumer online resale behavior measurement based on machine learning and BP neural network. *Journal of Intelligent & Fuzzy Systems*, *40*(2), 2121-2132.