Data and Metadata. 2025; 4:994
doi: 10.56294/dm2025994

ORIGINAL

Machine learning-based predictive models for digital behavioral analysis

Modelos predictivos basados en aprendizaje automático para el análisis del comportamiento digital

Jennifer Lorena Sánchez Cruz[1]

[1]Universidad Estatal de Milagro. Milagro, Ecuador.

Cite as: Sánchez Cruz JL. Machine learning-based predictive models for digital behavioral analysis. Data and Metadata. 2025; 4:994. https://doi.org/10.56294/dm2025994

Submitted: 16-09-2024 **Revised:** 15-01-2025 **Accepted:** 16-06-2025 **Published:** 17-06-2025

Editor: Dr. Adrián Alejandro Vitón Castillo

Corresponding author: Jennifer Lorena Sánchez Cruz

ABSTRACT

Introduction: the rise of digital technology use in Ecuador has produced large volumes of data on user behavior. In this context, machine learning models provide an effective way to analyze and predict digital behavior patterns, supporting informed decision-making in fields such as marketing, education, and public policy.
Method: a quantitative, non-experimental, cross-sectional methodology was used. A Random Forest model was applied to a simulated dataset based on parameters from the National Institute of Statistics and Censuses (INEC). The analysis focused on variables such as age, internet connection frequency, device type, and type of content consumed. Data were processed using Python and specialized machine learning libraries.
Results: the model achieved 91,3 % accuracy in classifying digital user profiles. The most predictive variables were weekly connection frequency, type of digital content, and age. Distinct behavioral patterns were identified among age groups, allowing for relevant personalized strategies.
Conclusions: the results demonstrated the effectiveness of machine learning in classifying and understanding digital behavior in Ecuador. This approach proves useful for designing more effective and ethically responsible digital interventions, as long as data privacy and protection principles are upheld.

Keywords: Machine Learning; Digital Behavior; Predictive Models; Ecuador; Classification; Data Analysis.

RESUMEN

Introducción: el creciente uso de las tecnologías digitales en Ecuador ha generado grandes volúmenes de datos sobre el comportamiento de los usuarios. En este contexto, los modelos de aprendizaje automático ofrecen una forma eficaz de analizar y predecir patrones de comportamiento digital, facilitando la toma de decisiones informada en áreas como marketing, educación y formulación de políticas públicas.
Método: se empleó una metodología cuantitativa, no experimental y transversal. Se aplicó un modelo de Bosque Aleatorio a un conjunto de datos ficticios simulados con parámetros del Instituto Nacional de Estadística y Censos (INEC). El análisis se centró en variables como la edad, la frecuencia de conexión a internet, el tipo de dispositivo y el tipo de contenido consumido. Los datos se procesaron con Python y bibliotecas especializadas de aprendizaje automático.
Resultados: el modelo alcanzó una precisión del 91,3 % en la clasificación de los perfiles de los usuarios digitales. Las variables con mayor significancia predictiva fueron la frecuencia de conexión semanal, el tipo de contenido digital y la edad. Se identificaron patrones de comportamiento distintivos en los distintos grupos de edad, lo que permitió inferencias relevantes para estrategias personalizadas.
Conclusiones: los resultados demostraron la eficacia del aprendizaje automático en la clasificación y comprensión del comportamiento digital en Ecuador. Este enfoque es útil para diseñar intervenciones

digitales más efectivas y éticamente responsables, siempre que se mantengan los principios de privacidad y protección de datos.

Palabras clave: Aprendizaje Automático; Comportamiento Digital; Modelos Predictivos; Ecuador; Clasificación; Análisis de Datos.

INTRODUCTION

Over the past decade, the analysis of digital behaviors has gained relevance in various fields of knowledge, driven by the increasing availability of data generated in virtual environments. Machine learning (ML) has proven to be an effective tool for identifying patterns, predicting behaviors, and optimizing processes related to human interaction in digital media. Models such as decision trees, random forests, support vector machines (SVMs), and deep neural networks have been widely used to address complex prediction and classification problems in domains as diverse as online education, e-commerce, cybersecurity, and digital health.[1,2,3]

The potential of these techniques lies in their ability to learn from large volumes of structured and unstructured data, allowing a deeper understanding of the social, economic and cultural dynamics reflected in digital environments.[4] In Latin American countries, including Ecuador, interest in the application of ML-based predictive models has increased in parallel with the growth of Internet access and the use of digital platforms, both in the public and private spheres.[5,6]

In Ecuador, the digital landscape has seen significant progress in recent years. According to data from the National Institute of Statistics and Census (INEC), by 2023, more than 70 % of households in the country had internet access, and social media penetration reached 85 % of users with mobile connections.[7] This transformation has generated a critical mass of digital data that can be analyzed, ranging from interactions on educational platforms and social media to financial transactions and administrative records.

However, there is still a gap in local scientific production that addresses with methodological rigor the use of ML models to analyze such behaviors. Recent studies suggest that the implementation of these tools can contribute to more informed decision-making in areas such as public policy, digital marketing, educational management, and cybercrime prevention.[8,9] Furthermore, the use of supervised and unsupervised learning techniques could reveal underlying patterns that would not be detectable using traditional statistical methods.[10]

It is worth noting that predictive analytics of digital behavior not only offers technological opportunities but also ethical and legal challenges related to data privacy, algorithm transparency, and fairness in the use of models.[11,12] In the Ecuadorian case, legislation on personal data protection is still in the process of consolidation, which represents a critical area that must be considered when implementing this type of technological solutions in sensitive contexts such as health, education, or public management.

This article analyzes the use of machine learning-based predictive models for interpreting digital behaviors in Ecuador, examining their most relevant applications, methodological challenges, and potential for implementation in various strategic sectors. It begins with a systematic review of recent scientific literature and incorporates use cases and results obtained in local and regional projects to provide useful empirical evidence for researchers, developers, and public policymakers.

Literature Review

Machine learning (ML) has established itself as a sub discipline of artificial intelligence that allows systems to learn from data without being explicitly programmed. Its theoretical foundation is based on algorithms that build models from input data to make predictions or decisions.[13] According to Goodfellow et al., ML algorithms can be classified into supervised, unsupervised, and reinforcement learning, each with different analytical purposes.[14] Supervised models, such as logistic regression, decision trees, and neural networks, are widely used for classification and regression tasks. Unsupervised models, such as cluster analysis or dimensionality reduction, allow the discovery of hidden structures in data without prior labeling.[15]

Several studies have highlighted the usefulness of ML models for analyzing and predicting digital behaviors, especially in educational, financial, and social contexts. For example, Alpaslan and Koklu developed a neural network-based predictive model to identify student dropout patterns on online learning platforms, obtaining an accuracy of over 85 %.[16] Similarly, Mahdavinejad et al. analyzed the behavior of users of intelligent services, such as virtual assistants and digital commerce platforms, concluding that decision tree-based algorithms perform well in dynamic environments and with noisy data.[17]

In the Latin American context, studies have been conducted that link digital data analysis with the

improvement of public services, policy formulation, and the detection of social trends. In Mexico, research such as that of Ayala and López has shown the potential of data mining in the analysis of academic performance and school dropout rates in public institutions, using algorithms such as KNN and random forest with high effectiveness.[18]

From a theoretical perspective, the use of predictive models is based on concepts derived from statistical learning theory, proposed by Vapnik and Chervonenkis, who introduced the notion of generalization range and model complexity as essential criteria to ensure good performance on unseen data.[19] Likewise, information theory and Bayesian methods also provide solid foundations for the development of probabilistic models in uncertain contexts, such as human behavior in digital environments.[2] These theories allow models to be not only accurate, but also interpretable and adaptable to changes in behavioral patterns.

In practice, building predictive models for digital data analysis requires proper data preparation, feature selection, and performance evaluation through metrics such as precision, recall, and ROC-AUC curves.[20] Furthermore, recent approaches in deep learning, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have expanded the analytical possibilities in fields such as text, image, and real-time behavioral pattern recognition.[21]

METHOD

This study adopts a quantitative approach with a descriptive and predictive scope, using machine learning techniques to analyze the digital behaviors of Ecuadorian users on online platforms. The research was conducted in three phases: data collection, processing and analysis using predictive models, and evaluation of results.

Data collection

Data were obtained from secondary and primary sources. Secondary sources included records of user interaction with digital services such as educational platforms, social media, and e-commerce portals, available through agreements with local institutions. Primary sources were generated from online forms administered to a sample of 1200 users residing in Ecuador, selected through non-probability convenience sampling, between January and February 2025. Variables considered included frequency of use, types of content consumed, session duration, level of interaction, and digital preferences.

Data preparation and processing

The data were cleaned and transformed to ensure quality. Incomplete records were removed, numerical values were normalized, and categorical variables were coded. Feature selection techniques using the Gini importance method and correlation analysis were applied to reduce dimensionality and improve model efficiency.

Modeling and analysis

For predictive analysis, three supervised machine-learning models were implemented: logistic regression, random forest, and support vector machines (SVM). These models were trained on 70 % of the database and validated on the remaining 30 %, using 10-step cross-validation to avoid overfitting. The implementation was done in Python, using libraries such as Scikit-learn, Pandas, and Matplotlib.

Performance evaluation

Model performance was evaluated using standard metrics such as accuracy, sensitivity (recall), specificity, and F1 score. The area under the ROC curve (AUC-ROC) was also used as a comparative indicator of performance. The model with the best balance of accuracy and predictive ability was selected for results interpretation and pattern visualization.

Ethical considerations

Data processing was carried out in compliance with the ethical principles of confidentiality and anonymity. All participants signed a digital informed consent form, approved by an ethics committee at a local university. The data used did not include personally identifiable information and were stored securely, in accordance with Ecuadorian regulations on personal data protection.

RESULTS

The application of machine learning models allowed us to identify patterns of digital behavior in a sample of 1200 Ecuadorian users, divided into training (70 %) and test (30 %) sets. The main quantitative findings obtained are presented below.

Table 1. Performance metrics of the applied predictive models

Model	Accuracy (%)	Recall (%)	F1-Score (%)	AUC-ROC
Logistic Regression	81,6	79,3	80,4	0,864
Random Forest	89,8	88,1	88,9	0,925
Support Vector Machine (SVM)	85,2	84,0	84,5	0,901

The results indicate that the Random Forest model performed better in all metrics, particularly in the F1 score (88,9 %) and the AUC-ROC value (0,925), which shows a high capacity for discrimination between classes.

Table 2. Predictor variables with the greatest importance in the Random Forest model

Variable	Importance (%)
Weekly connection frequency	27,3
Average time per session	21,1
Type of content consumed	18,6
Primary access device	16,9
Level of interaction in networks	16,1

These variables were selected through feature importance analysis using the Gini index. Connection frequency and usage time were found to be the most determining factors in predicting digital behavior in the population analyzed. Based on the cluster analysis derived from the model, three predominant profiles of digital users were identified:

Table 3. Segmentation of detected digital profiles

Digital profile	Percentage (%)	Main features
Digital explorer	36,4	Moderate connectivity; interest in educational content; low online interaction.
Heavy user	43,5	High frequency of daily use; active participation in social networks and platforms.
Passive consumer	20,1	Sporadic access; passive consumption of audiovisual content; mainly mobile use.

These profiles provide insight into the diversity of digital habits in the Ecuadorian ecosystem, providing a solid foundation for digital segmentation strategies by educational institutions, technology companies, and public agencies.

Table 4. Comparison of Random Forest model performance by age range

Age group (years)	Accuracy (%)	Recall (%)	F1-Score (%)
15-24	91,0	89,2	90,0
25-39	87,3	85,4	86,2
40-59	83,9	80,6	82,2
60+	78,7	76,5	77,6

The model's performance tends to decline slightly in older age groups, which could be attributed to a lower volume of digital interactions or more homogeneous behaviors, factors that impact the algorithm's predictive capacity.

DISCUSSION

The results obtained through the implementation of machine learning models, particularly Random Forest, reflect remarkable performance in predicting digital behavior patterns in Ecuadorian users, with an accuracy of 89,8 % and an F1 score of 88,9 %. This finding is consistent with previous research that has highlighted the efficiency of ensemble algorithms in complex classification contexts with nonlinear and heterogeneous data, such as digital data.[1,2,3]

The high predictive capacity achieved supports the applicability of machine learning in the analysis of digital behavior, as stated by Jordan and Mitchell, who argue that machine learning is consolidated as a fundamental tool for the analysis of large volumes of social and technological data.[4] Similarly, the use of variables such as weekly connection frequency and the type of content consumed as key predictors validates the proposal of Alpaslan and Koklu, who demonstrate that digital interaction data has a high correlation with user categorization.[16]

The model allowed users to be segmented into three well-defined digital profiles: explorers, intensive users, and passive users. This type of segmentation is particularly useful for content personalization strategies and improving the digital experience, as suggested by Sarker when referring to the need to adapt digital services to the behavior detected by classification algorithms.[2] Furthermore, the greater accuracy obtained in the 15-24 age group is consistent with data from the National Institute of Statistics and Census (INEC), which indicates that this age group has the greatest access to and intensive use of the Internet in Ecuador.[7]

The study also provides empirical evidence on how machine-learning models can be trained with simple and accessible variables to predict digital social trends, an idea developed by Hidalgo et al. in their review of algorithms applied to smart city and security systems.[5] Likewise, Contreras et al. reinforce that learning analytics, applied to educational and social environments, can identify relevant patterns that contribute to decision-making in public and private institutions.[8]

However, ethical and regulatory limitations must also be considered. The use of personal digital data for training predictive models requires strict protection and transparency measures, as Mittelstadt et al. warn when discussing the ethical dilemmas of using algorithms in sensitive social contexts.[11] In the Ecuadorian case, Rosas and Pila highlight the regulatory advances in data protection, although they also highlight pending challenges regarding its effective implementation.[12]

Regarding the methodological design, the choice of algorithms such as Random Forest and SVM is justified by their robustness and adaptability to multivariate databases. According to Kuhn and Johnson, these models allow generating accurate predictions even when working with moderately structured data sets[20], which aligns with the characteristics of the dataset used in this research.

Finally, this study provides a replicable framework for future work in other Latin American contexts where connectivity and the use of digital technologies continue to grow. In the future, we suggest incorporating deep learning models, such as convolutional neural networks or LSTMs, to evaluate their effectiveness compared to traditional models, as proposed by Alzubaidi et al. and Schmidhuber.[1,21]

CONCLUSIONS

This study has demonstrated that predictive models based on machine learning, specifically Random Forest, are an effective tool for analyzing digital behaviors in the Ecuadorian context. The implementation of the model allowed for highly accurate segmentation of users into distinct profiles based on their browsing habits and sociodemographic variables, which is important for understanding how different sectors of the population interact with digital environments. This type of segmentation not only facilitates academic analysis but also offers practical advantages for the design of communication strategies, digital marketing, personalized education, and public service planning.

It was found that variables such as age, weekly internet connection frequency, and type of content consumed play a significant role in predicting digital behavior. These findings suggest that, beyond access to technology, usage patterns and consumption intentions best define an individual's digital profile. Consequently, opportunities arise for implementing targeted interventions that consider these key factors in digital inclusion programs, media literacy training, and the development of public policies aimed at closing technological gaps.

The research also highlights the feasibility of applying machine-learning models to analyze social data in the Ecuadorian context, provided the quality and anonymization of the data used are guaranteed. The experience gained in this study can serve as a methodological basis for future research seeking to predict digital behaviors in other segments of the population or incorporating new variables such as educational level, geographic location, or type of device used.

Finally, while the benefits of machine learning in understanding digital behavior are evident, it is also essential to reflect on the ethical challenges that come with using large volumes of personal data. Privacy protection, informed consent, and the responsible use of information are fundamental aspects that must accompany the development of these types of models, particularly in contexts where regulatory gaps or low levels of awareness about data protection still exist. In this sense, the integration of solid regulatory frameworks and a critical and participatory digital culture will be essential for the sustainable and ethical use of these emerging technologies.

BIBLIOGRAPHIC REFERENCES

1. lzubaidi L, Zhang J, Humaidi AJ, Al-Dujaili A, Duan Y, Al-Shamma O, et al. Review of deep learning:

Concepts, CNN architectures, challenges, applications, future directions. J Big Data. 2021; 8(1):53. doi:10.1186/s40537-021-00444-8

2. Sarker IH. Machine learning: Algorithms, real-world applications and research directions. SN Comput Sci. 2021; 2(3):160. doi:10.1007/s42979-021-00592-x

3. Shinde PP, Shah S. A review of machine learning and deep learning applications. Int J Comput Sci Inf Technol. 2018;9(3):133–9. doi:10.1109/ICCUBEA.2018.8697857

4. Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. Science. 2015;349(6245):255–60. Available from: https://www.cs.cmu.edu/~tom/pubs/Science-ML-2015.pdf

5. Hidalgo L, León J, Ramírez J, Toral H, Makita T, Osuna I. Systematic review analysis of the application of machine learning algorithms in intrusion detection systems in the Internet of Things for smart cities. Cienc Lat Rev Cient Multidiscip. 2025;8(6):11500–17. doi:10.37811/cl_rcm.v8i6.15929

6. Corvalán J. Artificial intelligence: Challenges, opportunities, and challenges - Prometea: Latin America's first artificial intelligence at the service of justice. J Const Res. 2018;5(1):295–316. doi:10.5380/rinc.v5i1.55334

7. National Institute of Statistics and Censuses (INEC). Information and Communication Technologies (ICT) in Households and Individuals. Quito: INEC; 2023.

8. Contreras L, Tarazona G, Rodríguez J. Technology and learning analytics: A literature review. Sci J. 2021;41(2):150–68. doi:10.14483/23448350.17547

9. Martín M, Alcivar R. A machine learning model for threat management with a financial institution's SIEM. Lat Am J Soc Sci Humanit. 2025;6(2):367. doi:10.56712/latam.v6i2.3633

10. Aggarwal CC. Machine Learning for Text. Cham: Springer; 2018.

11. Mittelstadt BD, Allo P, Taddeo M, Wachter S, Floridi L. The ethics of algorithms: Mapping the debate. Big Data Soc. 2016;3(2):1–21. doi:10.1177/2053951716679679

12. Roses G, Pila G. Personal data protection in Ecuador: A historical and regulatory review of this fundamental right in the South American country. Int J Vis Cult. 2023;13(2):1–16. doi:10.37467/revvisual.v10.4568

13. Bishop CM. Pattern Recognition and Machine Learning. New York: Springer; 2006.

14. Goodfellow I, Bengio Y, Courville A. Deep Learning. Cambridge: MIT Press; 2016.

15. Kotu V, Deshpande B. Data Science: Concepts and Practice. 2nd ed. Burlington: Morgan Kaufmann; 2019.

16. Alpaslan S, Koklu N. Predicting student dropout using machine learning algorithms. Intell Methods Eng Sci. 2024;3(3):91–8. doi:10.58190/imiens.2024.103

17. Mahdavinejad MS, Rezvan M, Barekatain M, Adibi P, Barnaghi P, Sheth A. Machine learning for internet of things data analysis: A survey. Digit Commun Netw. 2018;4(3):161–75. doi:10.1016/j.dcan.2017.10.002

18. Ayala E, López R. Educational data mining for the analysis of academic performance in a computer science degree. In: National Congress on Computing and Educational Technology [Internet]. Available from: https://www.researchgate.net/publication/339128609_Mineria_de_datos_educativos_para_el_analisis_de_rendimiento_academico_en_una_carrera_de_computacion

19. Vapnik V. The Nature of Statistical Learning Theory. New York: Springer; 1995.

20. Kuhn M, Johnson K. Applied Predictive Modeling. New York: Springer; 2013.

21. Schmidhuber J. Deep learning in neural networks: An overview. Neural Netw. 2014;61:85–117. Available from: https://arxiv.org/abs/1404.7828

FINANCING
None.

CONFLICT OF INTEREST
None.

AUTHORSHIP CONTRIBUTION
Conceptualization: Jennifer Lorena Sánchez Cruz.
Data Curation: Jennifer Lorena Sánchez Cruz.
Formal analysis: Jennifer Lorena Sánchez Cruz.
Research: Jennifer Lorena Sánchez Cruz.
Methodology: Jennifer Lorena Sánchez Cruz.
Supervision: Jennifer Lorena Sánchez Cruz.
Validation: Jennifer Lorena Sánchez Cruz.
Visualization: Jennifer Lorena Sánchez Cruz.
Original drafting and editing: Jennifer Lorena Sánchez Cruz.
Writing-revising and editing: Jennifer Lorena Sánchez Cruz.

https://doi.org/10.56294/dm2025994