IJSES
Volume 10, Issue 3, pp. 16-20, 2026.
International Journal of Scientific Engineering and Science
ISSN (Online): 2456-7361

Applying Machine Learning Models to Classify User Behavior Through Digital Footprint Data

Trinh Thi Huong[1], Le Ngoc Anh[2], Mai Ngoc Minh[3], Tran Minh Khue[3], Trinh Lan Phuong[3], Le Thuc Anh[3]
[1]Faculty of Mathematical Economics, Thuongmai University, Hanoi, Vietnam
[2] K60V1, Faculty of Mathematical Economics, Thuongmai University, Hanoi, Vietnam
[3]K60V2, Faculty of Mathematical Economics, Thuongmai University, Hanoi, Vietnam

Abstract*—Digital footprint data from smartphones provide new opportunities for understanding and classifying user behavior. This study evaluates machine learning models for identifying daily activities using spatial–temporal and movement-related features from the publicly available MyDigital Footprint dataset, collected from 31 volunteers over two months. Key variables include location, speed, temporal indicators, and cyclical time encoding. Gaussian Naive Bayes, Decision Tree, and Random Forest models are implemented and assessed using a stratified 80/20 split. Results show that tree-based models outperform the probabilistic baseline, with Decision Tree and Random Forest achieving accuracies of 0.98 and 0.96, respectively. The findings highlight the effectiveness of non-linear models in capturing complex behavioral patterns*.

***Keywords*—** *Machine Learning Models, User Behavior, Digital Footprint Data*.

I. INTRODUCTION

In recent years, digital transformation has become an inevitable trend, closely tied to the explosive growth of data and socio-economic activities in the digital environment. The widespread adoption of the Internet, mobile devices, digital platforms, and technologies such as artificial intelligence and cloud computing has fundamentally changed how people work, communicate, consume, and conduct business. Activities ranging from e-commerce and digital finance to online marketing and business management generate enormous, diverse, and continuously evolving amounts of data. In this context, digital data plays a crucial role in the activities of individuals, businesses, organizations, and governments. All activities of individuals and organizations are recorded in the digital environment, including digital footprint data, which consists of information obtained from digitally traceable behavior and online presence. With the rapid development of machine learning algorithms, digital footprint data is increasingly utilized across various fields, including economics, business, and social management. Among these applications, the use of machine learning to classify user behavior has garnered significant attention from researchers and organizations. However, given the complexity of data and user behavior, further research on this topic is essential.

II. RESEARCH OVERVIEW

Digital footprint data encompasses the comprehensive set of information generated during an individual's interactions with digital devices, online platforms, and modern technology ecosystems [1]. These footprints are continuously created as users access websites, utilize social networks, install and operate mobile applications, conduct online searches, engage in online shopping, send emails, or navigate with devices equipped with location services and sensors.

Globally, the analysis of digital footprint data is applied across various economic and business sectors. In China, for instance, the travel itineraries of tourists are analyzed to understand tourism behavior, which the government subsequently uses to develop targeted tourism development strategies [2]. Additionally, businesses leverage digital footprint data to analyze customer profiles and behavior, enabling them to devise marketing strategies tailored to their target customer segments and enhance their competitive advantage [3]. In the finance and banking sector, research indicates that users are more inclined to share digital footprint data [4]. Moreover, individuals are more likely to share their data with academic institutions for research purposes than with private companies or government entities.

In Vietnam, digital footprint data and artificial intelligence models are being utilized to create personalized learning experiences, particularly through models that track and predict learners' knowledge states over time [5]. Research demonstrates that machine learning models possess strong predictive capabilities for user behavior and have potential applications in specific contexts, such as the education sector. In the realm of online sales, data from the online shopping journey can forecast customer touchpoints [6] by analyzing online shopping information from customers in the tourism industry. By applying the K-Means algorithm for clustering, the authors identified target customer segments and employed various trained machine learning models to predict purchase decisions. This research enhances the understanding of customer journeys by integrating recommendation and decision-making systems, providing a practical predictive model applicable across various business types.

In 2021, Campana and Delmastro published a dataset encompassing smartphone sensors, physical proximity information, and social media data. The tracked variables include location (latitude/longitude), smartphone sensors (velocity, acceleration), Wi-Fi information, and interactions on

online social networks (ONS) [7], [8]. The authors collected real-world data from the mobile devices of 31 volunteers over two months and presented machine learning applications such as social relationship analysis, daily activity recognition, and contextual recommendation systems. This meticulously constructed and open-source digital footprint dataset is utilized by the authors and other researchers in contextual recognition studies and user behavior modeling.

III. Research Methodology and Proposed Model

3.1 Research methodology

The classification problem involves examining the relationship between categorical dependent variables and independent variables, which can be continuous or discrete. The classification problem is stated as follows: consider the stochastic dependency relationship between input and output, described by the joint distribution $F_{(x,y)}(.)$. Then, for an input vector x , the output variable $y \in Y = \{c_1, c_2, \ldots, c_K\}$ will take one of K different classes according to a conditional distribution.

Loss Function

A classifier is a specific instance of a learning algorithm, whereby for a given input x, the algorithm returns an estimate $\hat{y} = \hat{c} = h(x, \alpha)$, taking a value in the set $Y = \{c_1, c_2, \ldots, c_K\}$. Then, the loss function is defined as follows:

$$L(c, \hat{c}) = I(c = \hat{c}) = \begin{cases} 0 & if \quad c = \hat{c} \\ 1 & if \quad c \neq \hat{c}. \end{cases}$$

The notation $I(.)$ is often called the criterion function. In this case, when considering the loss matrix, it will have the form: on the main diagonal, it is 0 (i.e., no loss when the classification is correct), and all elements outside the main diagonal are 1.

The goal of the classification problem for a given x) is to find a predictor $\hat{y} = \hat{c}(\mathbf{x}) = h(x, \alpha)$ that minimizes the objective function

$$\sum L_{(c_k, \hat{c}(\mathbf{x}))} \, P\{ y = c_k | x \}.$$

The objective function above can be considered as the average of the loss function with weights equal to the conditional probability of the random variable $y = c_k$ given x .

Common estimation methods:

- The GaussianNB (NB) model is a Naive Bayes classifier provided by the scikit-learn library. This classifier is suitable for continuous feature variables and assumes that each feature follows a normal (Gaussian) distribution. The algorithm also relies on the assumption that features are conditionally independent of class labels [9].
- The Decision tree (DT) classification model is used as a supervised classification method, provided by the scikit-learn library [10]. Decision tree models are suitable for data containing both continuous and discrete variables, do not require assumptions about the probability distribution of the input data, and are capable of modeling non-linear relationships between features. Decision tree algorithms operate on the principle of dividing the feature space into sub-regions through a series of conditional tests. At each node, the model selects features and a threshold for division so that the data's disorder is minimized. In classification problems, the criteria commonly used are Gini impurity or entropy.
- Random forests (RF) are classification models belonging to the ensemble learning group, provided by the scikit-learn library [10]. Random forests work by constructing multiple independent decision trees on random subsets of data and features, then aggregating the prediction results (majority voting) to arrive at a final label.

3.2. Data

The study used the secondary MDF (MyDigitalFootprint) dataset, publicly available at: https://github.com/contextkit/MyDigitalFootprint. The MDF data was collected from the personal smartphones of 31 volunteers (6 women and 25 men) who participated for two months in three cities in the Tuscany region of Italy [7]. To collect the MDF data, project members met with the volunteers and explained the details of the information to be collected. The data collection period was 2 months (60 days), and volunteers could optionally turn off tracking during the study. Of the 31 volunteers, 12 participated in the project for less than 15 days, while the remaining volunteers participated for 15 to 45 days.

MDF data includes the following six main groups of information:

- User Activity: Collects information about each user's activity, including:
- Location: Information about the user's geographical location, including latitude, longitude, and exact location.
- User Activity (Android Activity Recognition): User activity is identified through Android's Activity Recognition system. The system can recognize human gait (e.g., running and walking) and modes of locomotion (e.g., sitting in a car and riding a bicycle).
- Phone Status: Includes information about system sound settings, battery information, screen status, and physical sensor data that can describe the situation the user is in. For example, the ringtone and screen might be off because the user is in a business meeting.
- Mobile Network: Includes information related to mobile networks, contacts, and phone calls.
- Wireless Interface: Wireless communication interfaces such as Wi-Fi and Bluetooth. Specifically, MDF collects information about Bluetooth connectivity, Wi-Fi access, and a list of Wi-Fi hotspots (Wi-Fi P2P) near the user's phone.
- User Preferences: Includes information about the applications the user uses during the day and the time spent using those applications.
- External Information: Information not directly related to the mobile device that can be used to refine the user's context, such as weather.

This study uses the following variables.

International Journal of Scientific Engineering and Science
Volume 10, Issue 3, pp. 16-20, 2026. ISSN (Online): 2456-7361

TABLE 1: Description of observed variables

No	Variables	Meaning	Value
1	*user_id*	Volunteer ID	1, 2, …, 31
2	*hour_1_24*	Hour	1, 2, …, 24
3	*minute*	Minute	1, 2, …, 60
4	*dayofweek*	Day of week	0,1,…,6
5	*is_weekend*	Binary variable for weekend	0: Day of week 1: Weekend
6	*label*	User activities	Home, Free Time, School, Workplace, External School, and Holiday
7	*time*	Time (Unix timestamp, ms)	
8	*lat, long, acc, alt*	latitude, longitude, accuracy and Altitude	
9	*speed*	User movement speed	[0, 52.8]
10	*bearing*	*Bearing*	
11	*postime*	Time of location recording (timestamp)	
12	*sin_hour*	Sinusoidal variable of the time of day (processing a 24-hour cycle)	[-1, 1]
13	*cos_hour*	Cosine transformation of hour (24-hour cycle encoding) (processing a 24-hour cycle).	[-1; 1]

3.3. Proposed research model

This study examines the influence of spatial, temporal, movement characteristics, and periodic factors on classifying user behavior. Specifically, it classifies user activities with labels such as Home, Free Time, School, Workplace, External School, and Holiday. In this study, we use a pooled dataset of all users. A similar approach could be applied to groups of users or individual users.

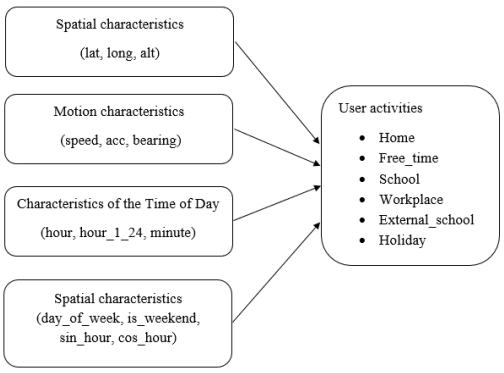

Fig.1: Research model framework for classifying user activity based on spatial and temporal data.

IV. RESEARCH RESULTS

4.1 Descriptive statistics of the dataset

Analysis of data from 31 participants reveals a notable discrepancy in user participation and frequency of data recording (Fig. 2). Some individuals contribute thousands of records, while others provide only a few dozen to a few hundred. This variation is due to differences in device usage duration, levels of data sharing, or individual lifestyle habits. Moreover, this disparity suggests an imbalance within the dataset, with a small number of individuals supplying the majority of the data.

The chart illustrating average movement speed per user indicates significant variations in movement behavior among participants. Some users exhibit very low average speeds, suggesting they remain primarily stationary or operate within a limited range. In contrast, others demonstrate high average speeds, indicative of frequent movement patterns or the use of faster transportation methods. These differences highlight the influence of personal factors on the formation of an individual's digital footprint.

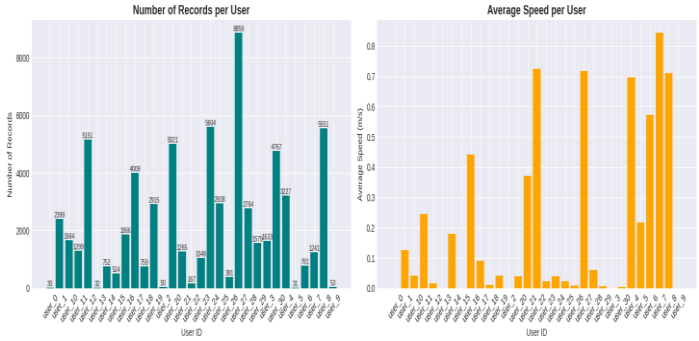

Fig.1: Distribution graph showing the number of records and average speed per user.

4.2. Results of user behavior classification

The entire dataset, after preprocessing, will be divided into two parts: a training set and a test set, with an 80/20 ratio. A label-based stratification strategy will be employed to maintain the proportion of active classes in both datasets.

Table 2 shows that the Naive Bayes model (NB), when trained on the complete dataset, achieved an accuracy of 0.71 and a weighted F1-score of 0.70. These results indicate that the model demonstrated average to good classification performance across the entire dataset. However, a detailed analysis reveals significant differences in classification performance among active classes, highlighting the heterogeneity and varying degrees of behavioral overlap between classes in the dataset.

Considering each activity class individually, the "Home" class achieved high performance with a recall value of 0.94 and an F1-score of 0.80. This indicates that the model effectively identifies samples belonging to the Home" class, likely due to the large sample size of 6,511 and the relatively stable spatial and temporal characteristics, which facilitate the learning of classification rules.

The "Holiday" class also performed well, with an F1-score of 0.83, reflecting its clear ability to distinguish itself from other activity classes. However, there are instances where the "Holiday" class can be confused with the "Home" class, suggesting some similarity in the living contexts of the two categories.

For the "Free Time" class, the model achieved high precision (0.89) but low recall (0.48). This indicates that the model predicts the label "Free Time" only in cases of high certainty, resulting in a significant number of actual samples belonging to this class being missed. This issue reflects the overlapping characteristics between "Free Time" and other

classes, such as "Home" or "Workplace" which blur the classification boundaries.

The "School" and “Workplace" classes exhibited lower classification performance, with F1-scores of 0.41 and 0.50, respectively. This suggests that the model struggles to differentiate between learning and working activities and other activities, likely due to similarities in time and location context, especially when these activities occur at home or in outdoor settings.

TABLE 2: Table of results for evaluating the classification model

Label	Precision			Recall			F1-score			Support		
	NB	DT	RF	NB	DT	RF	NB	DT	RF	NB	DT	RF
External School	0.14	0.76	0.87	0.95	0.90	0.81	0.25	0.82	0.84	81	81	81
Free Time	0.89	0.98	0.95	0.48	0.98	0.94	0.62	0.98	0.95	4847	4847	4847
Holiday	0.96	0.98	1.00	0.73	1.00	0.99	0.83	0.99	0.99	673	673	673
Home	0.70	0.99	0.96	0.94	0.99	0.97	0.80	0.99	0.97	6511	6511	6511
School	0.48	1.00	1.00	0.36	0.99	0.92	0.41	0.99	0.96	631	631	631
Workplace	0.52	0.98	0.98	0.49	0.97	0.97	0.50	0.98	0.98	933	933	933
Accuracy							0.71	0.98	0.96	13676	13676	13676
Macro avg	0.61	0.95	0.96	0.66	0.97	0.94	0.57	0.96	0.95	13676	13676	13676
Weighted avg	0.76	0.98	0.96	0.71	0.98	0.96	0.70	0.98	0.96	13676	13676	13676

Notably, the " External School" class has a very high recall value (0.95) but a low precision (0.14), indicating that the model tends to overpredict this class, leading to many incorrect predictions. This issue is likely a result of severe data imbalance, as the "External School" class contains only 81 samples, which hinders the model's ability to establish clear boundaries for this activity class.

Table 2 demonstrates that the Decision Tree model, when trained on the entire dataset, achieved exceptionally high classification performance on the test set. Specifically, the model attained an accuracy of approximately 0.98, alongside a precision of 0.95, a recall of 0.97, and an average F1-score (macro) of 0.96. Additionally, the weighted averages for precision, recall, and F1-score all approached 0.98, indicating effective performance even in the presence of significant imbalances between activity classes. In comparison to the baseline model and the Naive Bayes model, the Decision Tree exhibited clear superiority in classification performance. This outcome illustrates the model's capability to effectively leverage non-linear relationships between spatial and temporal features, significantly enhancing its ability to differentiate activity classes with overlapping characteristics.

When considering each activity class individually, the model achieved outstanding performance for the "Home" class, with a precision of 0.99, a recall of 0.99, and an F1-score of 0.99 across 6,511 samples. This result indicates that the model accurately identifies activities occurring at home, benefiting from the large sample size and stable behavioral features.

The analysis of the Random Forest model revealed an accuracy of approximately 0.959 and a weighted F1-score of around 0.959. Compared to the baseline model, which consistently predicted the "Home" label with an accuracy of about 0.476, the Random Forest improved accuracy by nearly 48 percentage points, demonstrating that the temporal and spatial features provided significant discriminatory information that the model effectively utilized.

For each class, the majority of labels achieved an F1-score of 0.95 or higher. The "Home" class, which had the largest sample size in the test set, achieved an F1-score of approximately 0.97, reflecting stable and consistent identification across all users. The "Free Time" class attained an F1-score of approximately 0.95, indicating that the model managed the overlap between "Free Time" and "Home" effectively, despite these activities often occurring in the same location. The "Holiday" class reached an F1-score of nearly 0.99, demonstrating a high degree of separation from other activities. The "School" class achieved an F1-score of approximately 0.96, while the "Workplace" class achieved an F1-score of about 0.98, indicating that these two daytime activities are relatively well distinguished. Although the rare " External School" class had a very limited sample size, it still achieved an F1-score of approximately 0.84. While this class had the lowest performance, an F1-score above 0.8 is still considered acceptable given the strong data imbalance and small number of observations.

Table 2 also shows that the decision tree model yielded the best overall estimation results.

V. Conclusion and Recommendation

After conducting practical research and experimental implementation on the MyDigitalFootprint dataset, the project achieved its objectives in developing a process for classifying user behavior using machine learning models. The analysis of data from 31 users, encompassing spatial and temporal characteristics, demonstrated that human behavior exhibits structured spatial–temporal regularities; it consistently follows cyclical and distinct geographical patterns. Supervised machine learning algorithms have proven to be effective tools for exploiting these patterns, enabling the automatic and accurate identification of living contexts.

The experimental results indicate that tree-based models, such as Decision Trees and Random Forests, outperform probabilistic Naive Bayes models. Naive Bayes assumes conditional independence among features, which may not hold in this dataset. In contrast, Decision Trees and Random Forests have shown the ability to delineate complex behavioral spaces, achieving optimal accuracy and F1-scores. Notably, processing raw data through technical steps, such as time-cycle coding using trigonometric functions (sin/cos hour), has enabled the models to capture the continuity of

circadian rhythms, significantly enhancing accuracy in distinguishing activity states over time.

Based on the results and existing limitations, the research team proposes several directions for future improvement and development. First, regarding practical application, it is recommended to establish an adaptive learning mechanism. The system should begin with a general model to provide immediate service to users. Then, through interaction and the accumulation of real-world data, the system will automatically evolve or merge with a personalized model to refine accuracy over time. From a technical data processing perspective, deeper interventions are required for minority label classes. Implementing algorithms to rebalance datasets or utilizing data augmentation techniques will increase the model's sensitivity to less repetitive but significant behaviors in users' lives. Additionally, integrating supplementary sensor data sources, such as network connection status, ambient light intensity, or application activity levels, will enrich the behavioral context and reduce reliance solely on GPS coordinates, which can be inaccurate due to environmental conditions.

In the future, research should extend to deep learning models capable of long-term memory, such as LSTM or attention models (Transformers). These architectures are promising for better leveraging the sequential nature of behavior—where present behavior is often closely related to that of preceding moments. Equally important is addressing security and privacy concerns. The research team recommends developing a federated learning approach, which allows the model to be trained directly on the user's device without transmitting sensitive location data to a server, thereby ensuring maximum privacy while maintaining high classification performance.

Finally, expanding the data collection scope to include more diverse user groups in terms of occupation, age, and geographic area will help validate the sustainability of the proposed algorithms. This not only enhances the scientific value of the research but also establishes a solid foundation for deploying behavior recognition solutions in practical fields such as proactive healthcare, smart city management, and personalized utility services on mobile devices.

REFERENCES

[1] Grant Longstaff, “What is a digital footprint?,” *The University of Law*, 2025.

[2] N. Mou, Y. Zheng, T. Makkonen, T. Yang, J. J. Tang, and Y. Song, “Tourists’ digital footprint: The spatial patterns of tourist flows in Qingdao, China,” *Tour. Manag.*, vol. 81, p. 104151, 2020.

[3] Y.-J. Chen, Y.-M. Chen, Y.-J. Hsu, and J.-H. Wu, “Predicting consumers’ decision-making styles by analyzing digital footprints on facebook,” *Int. J. Inf. Technol. \& Decis. Mak.*, vol. 18, no. 02, pp. 601–627, 2019.

[4] R. McDonald, A. Skatova, and C. Maple, “Attitudes towards sharing digital footprint data: a discrete choice experiment,” *Int. J. Popul. Data Sci.*, vol. 8, no. 3, 2023.

[5] Phạm Hoàng An, “Mô hình truy vết tri thức người học kết hợp động lực thời gian: Phân tích khả năng dự báo từ Saint và dữ liệu EDNET,” *Tạp chí Khoa học Trường Đại học Mở Hà Nội*, p. 549, 2025.

[6] Lâm Thị Bích Ngân and Thái Kim Phụng, “Xây dựng mô hình dự đoán điểm chạm của khách hàng dựa trên dữ liệu hành trình mua sắm trực tuyến,” *Tạp chí Nghiên cứu Tài chính-Marketing*, vol. 16, no. 6, pp. 42–54, 2026.

[7] M. G. Campana and F. Delmastro, “MyDigitalFootprint: An extensive context dataset for pervasive computing applications at the edge,” *Pervasive Mob. Comput.*, vol. 70, p. 101309, 2021.

[8] M. G. Campana and F. Delmastro, “On-device modeling of user’s social context and familiar places from smartphone-embedded sensor data,” *J. Netw. Comput. Appl.*, vol. 205, p. 103438, 2022.

[9] S. Raschka and V. Mirjalili, *Python machine learning: Machine learning* and *deep learning with Python, scikit-learn, and TensorFlow 2*. Packt publishing ltd, 2019.

[10] J. E. Black, J. K. Kueper, and T. S. Williamson, “An introduction to machine learning for classification and prediction,” *Fam. Pract.*, vol. 40, no. 1, pp. 200–204, 2023.