THE SCIENTIFIC TEMPER (2025) Vol. 16 (12): 5157-5165

E-ISSN: 2231-6396, ISSN: 0976-8653

Doi: 10.58414/SCIENTIFICTEMPER.2025.16.12.02

https://scientifictemper.com/

RESEARCH ARTICLE

Label-Aware Imputation with Cluster Refinement for Smartphone Usage Analytics in Educational Institutions

Vimala S[1*], G. Arockia Sahaya Sheela[2]

Abstract

The accurate handling of missing values remains a crucial step in data preprocessing, particularly in behavioral analytics where data incompleteness can distort pattern recognition and predictive modeling. This study presents a novel Label-Aware Imputation with Cluster Refinement (LAICR) framework designed specifically for smartphone usage datasets collected from educational institutions. The method partitions the dataset by usage-level labels (Low, Moderate, High), applies class-specific imputation using iterative reconstruction for numerical data and mode-based filling for categorical data, and refines results through K-Means clustering to improve local consistency.Experiments conducted on school and college datasets demonstrate significant improvements over standard global imputation techniques. The proposed method achieved an RMSE of 0.4575 and R^2 of 0.7735 for the school dataset, and RMSE of 0.4876 and R^2 of 0.7636 for the college dataset, outperforming global iterative and statistical baselines. Additionally, classification performance on imputed datasets reached 99.3% accuracy with XGBoost, indicating strong preservation of feature discriminability.The novelty of this work lies in combining label-awareness with intra-class cluster refinement, effectively reducing reconstruction error and preserving behavioral structure. This approach enhances the reliability of smartphone usage analytics, enabling more robust predictive modeling and behavioral interpretation in educational contexts.

Keywords: Smartphone usage, Academic performance, Missing imputation, Machine learning, Clustering.

Introduction

The increasing penetration of smartphones in everyday life has profoundly transformed how students engage with learning, communication, and recreation (Ben Hkoma *et al*., 2025). Among young populations, particularly school and college students, smartphones have become indispensable tools for accessing educational content, interacting on social platforms, and participating in online communities (Idoiaga *et al*., 2025). While these devices bring many benefits, including instant access to information, opportunities for collaborative learning, and tailored educational experiences, they also pose issues like overuse, dependence on technology, and increased digital distractions (Tang *et al*., 2025). Understanding these usage patterns has become an important focus of recent educational, psychological, and computational research, as it provides insights into students' well-being, academic engagement, and cognitive development (Du & Wang 2025).

To investigate such behavioral dynamics, data-driven approaches have gained traction, wherein structured smartphone usage surveys and digital activity logs are collected from student populations (Vimala 2025). These datasets typically contain both categorical information (e.g., gender, daily usage pattern, app categories, activity type) and numerical variables (e.g., screen time duration, number of app launches, call duration). However, the data collection process itself is often prone to incompleteness. Students may choose not to disclose certain details due to privacy concerns, lack of awareness, or discomfort in reporting sensitive usage information. Additionally,

[1]PhD Scholar, Department of Computer Science, St. Joseph'sCollege(Autonomous), Tiruchirappalli -2, Affiliated to Bharathidasan University, Tamil Nadu, India.

2Assistant Professor, Department of Computer Science, St. Joseph'sCollege(Autonomous), Tiruchirappalli -2, Affiliated to Bharathidasan University, Tamil Nadu, India.

***Corresponding Author:** Vimala S, PhD Scholar, Department of Computer Science, St. Joseph'sCollege(Autonomous), Tiruchirappalli -2, Affiliated to Bharathidasan University, Tamil Nadu, India, E-Mail: vimalas_phdcs@mail.sjctni.edu

How to cite this article: Vimala, S., Sheela, G.A.S. (2025). Label-Aware Imputation with Cluster Refinement for Smartphone Usage Analytics in Educational Institutions. The Scientific Temper, **16**(12):5157-5165.

Doi: 10.58414/SCIENTIFICTEMPER.2025.16.12.02

Source of support: Nil

Conflict of interest: None.

Received: 04/11/2025 **Accepted:** 25/11/2025 **Published:** 25/12/2025

technical issues such as skipped questions in digital forms or poorly synchronized logging mechanisms contribute further to missing entries. As a result, smartphone usage datasets often contain varying degrees of missing values, which can significantly affect subsequent analysis, modeling, and interpretation (Vimala & Sheela 2025).

Missing values in behavioral datasets are particularly problematic because they do not occur uniformly at random (Fontaine *et al*., 2025). For example, students with high smartphone usage levels may be more reluctant to disclose sensitive information such as nighttime usage or social media activity. This creates biased missingness patterns, making simple imputation approaches inadequate. Traditional imputation techniques, including global mean, median, or mode substitution, frequently do not adequately represent the complexity of behavioural patterns (Lagiou *et al*., 2025). They impose global averages on heterogeneous groups, which can distort the distribution of the data, reduce the separability between behavioral categories, and degrade the performance of predictive models (Schumann *et al*., 2025). Moreover, missing categorical values, which are common in behavioral datasets, are especially challenging to impute because they lack inherent numerical relationships that standard statistical methods rely upon.

The motivation behind this study stems from the need to develop a more robust and behaviorally consistent imputation framework that addresses these limitations. In the context of educational settings, accurate reconstruction of missing data is essential not only for improving the quality of predictive models but also for preserving behavioral authenticity. Behavioral labelssuch as "Low," "Moderate," and "High" smartphone usage levelscarry significant contextual meaning. If imputation ignores these class distinctions, it risks blending distinct usage behaviors into misleading averages. Therefore, a method that leverages class labels during imputation can help maintain the structural integrity of the data while improving accuracy.

The problem can be defined as follows: given a smartphone usage dataset with a mix of categorical and numerical features and non-random missing values, how can missing entries be accurately imputed while preserving the behavioral patterns of different usage groups? Existing global methods fail to adapt to label-specific variations, and advanced machine learning imputers often treat the dataset as homogeneous, overlooking behavioral class boundaries. This leads to higher reconstruction errors and reduced classification performance in downstream tasks such as smartphone usage prediction or risk-level identification.

To address this, the objective of this study is to design and evaluate a Label-Aware Imputation with Cluster Refinement (LAICR) framework. The framework partitions the dataset based on smartphone usage labels and applies class-wise imputation using iterative numerical reconstruction for continuous variables and mode imputation for categorical variables. Subsequently, K-Means clustering is employed within each label partition to refine imputed values based on local cluster centroids, ensuring greater coherence and minimizing intra-class variance. This approach aims to deliver both high reconstruction accuracy and behavioral fidelity.

The significance of this research lies in its potential to enhance the reliability of smartphone usage analytics in educational settings. Accurate imputation not only supports more trustworthy behavioral modeling but also facilitates effective intervention strategies for identifying and mitigating problematic usage patterns among students. Furthermore, the proposed framework integrates well with downstream classification models, as evidenced by the high predictive accuracy achieved after imputation. By preserving both numerical and categorical structures within each behavioral segment, this approach contributes a novel, domain-adapted solution to a fundamental challenge in educational data science. Ultimately, the proposed method supports better decision-making in academic institutions, enabling targeted mental health and digital wellness initiatives based on more reliable data.

Literature Review

Imputation of missing values plays a fundamental role in ensuring data integrity, model reliability, and reproducibility in data-driven research. Missing data can arise from multiple causes, such as user non-responses, measurement errors, privacy restrictions, or data collection failures. In recent years, significant advancements have been made in imputation strategies, particularly for structured numeric data and unstructured text data. The choice of imputation method has direct implications for the accuracy of downstream machine learning models, making it an essential step in modern data preprocessing pipelines.

For numeric data, early imputation strategies primarily focused on simple statistical methods, including mean, median, and mode imputation. These methods are computationally efficient and easy to implement, making them a common choice for baseline experiments. However, their major limitation lies in the inability to reflect the underlying data distribution. By replacing missing values with central tendencies, these methods often underestimate variance and introduce bias into the dataset (Prakash *et al*., 2024; Aljuaid & Sasi, 2016). Consequently, such approaches may distort relationships between variables, affecting both descriptive and inferential analyses.

To overcome these limitations, machine learning–based imputation techniques such as k-Nearest Neighbors (kNN) and Iterative Imputer have gained popularity. These methods exploit correlations and local data structures to predict missing values more accurately. kNN imputation identifies similar samples based on distance metrics and imputes missing values from neighboring instances. Iterative

imputation, often implemented through Bayesian Ridge regression, models each variable with missing values as a function of other variables in multiple iterations. Although these methods produce more accurate imputations than simple statistical techniques, they demand higher computational resources and careful parameter tuning (Prakash *et al*., 2024; Tsai *et al*., 2018).

Beyond deterministic models, advanced statistical approaches such as Gaussian Copula Imputation have shown strong performance in various scenarios. By modeling the joint distribution of all variables, Gaussian Copula effectively handles complex dependency structures and provides more reliable imputations under different missing data mechanisms. This probabilistic foundation makes it particularly robust when dealing with multivariate and partially observed datasets (Prakash *et al*., 2024).

Another promising category involves Class Center-Based Missing Value Imputation (CCMVI). Unlike purely global methods, CCMVI uses class-level centroids to guide the imputation process. By leveraging the structure of labeled data, it offers improved accuracy, particularly for mixed-type datasets where categorical and numeric variables interact (Tsai *et al*., 2018). Class-based strategies also preserve intra-class variance more effectively than global approaches, which can be critical in behavioral and health-related studies.

Recently, diffusion models have emerged as powerful tools for imputation tasks. Methods such as Conditional Score-Based Diffusion Models for Tabular Data (TabCSDI) can model complex, non-linear distributions and handle both numeric and categorical variables simultaneously. These models iteratively reconstruct missing entries by learning score functions in the data manifold, leading to highly accurate imputations in sparse and heterogeneous datasets (Zheng & Charoenphakdee, 2022). The ability of diffusion models to generate coherent and distributionally faithful imputations marks a substantial leap beyond classical techniques.

For text data, imputation presents additional challenges due to its unstructured nature. Deep learning models, particularly those leveraging n-gram representations and language modeling, have been adapted to fill missing text segments. These models excel in multilingual and large-scale datasets, providing context-aware imputations that preserve semantic meaning (Biessmann *et al*., 2018). However, their computational cost can be significant, especially when dealing with transformer-based architectures in real-world applications.

Interestingly, linear n-gram models have shown competitive performance with deep learning approaches at a lower computational cost. These models capture local linguistic structures effectively and are more interpretable, making them suitable for applications where efficiency is prioritized (Biessmann *et al*., 2018). Furthermore, fuzzy and clustering-based approaches have gained traction as hybrid solutions, combining fuzzy logic with regression or clustering to capture latent text patterns. Such models handle ambiguity and uncertainty in text data more gracefully (Bridge-Nduwimana *et al*., 2025).

The selection of an imputation method depends on several factors: the type of missingness mechanism (MCAR, MAR, or MNAR), the data modality (numeric, categorical, or text), the complexity of dependencies among variables, and computational constraints (Aljuaid & Sasi, 2016; Sivakani *et al*., 2025). Simple methods remain relevant in low-resource settings, whereas advanced statistical and machine learning approaches are more appropriate for high-dimensional, complex datasets. Diffusion-based and hybrid models represent the frontier of research, offering the best balance between accuracy and adaptability across data types.

Finally, mean and mode imputation are widely used for their simplicity, they are best suited for exploratory analysis rather than high-stakes modeling. Gaussian Copula and class-center approaches improve reliability in structured data, whereas diffusion models offer state-of-the-art performance in challenging scenarios. For text, linear n-gram and deep learning models offer flexible solutions depending on resource availability. The integration of multiple imputation strategies into hybrid frameworks represents a growing trend, aiming to achieve both efficiency and accuracy in real-world applications (Ahmad *et al*., 2024).

Methodology

The proposed research introduces a Label-Aware Imputation with Cluster Refinement (LAICR) framework designed to address the challenge of missing data in smartphone usage survey datasets. The methodology follows a structured pipeline beginning with preprocessing, progressing through label-wise imputation and refinement, and concluding with dataset reconstruction for downstream classification tasks.

Overview

The overall workflow in Figure 1 begins with raw data ingestion, followed by preprocessing that includes data type identification, feature encoding, and initial missing value handling. The processed dataset is partitioned based on smartphone usage labels (Low, Moderate, High). Within each partition, imputation is performed separately using iterative methods for numerical variables and mode filling for categorical variables. To enhance local accuracy, K-Means clustering is applied inside each class partition, refining the imputed values based on cluster statistics. The refined partitions are then recombined to form the final imputed dataset, which is subsequently used for classification model training and evaluation.

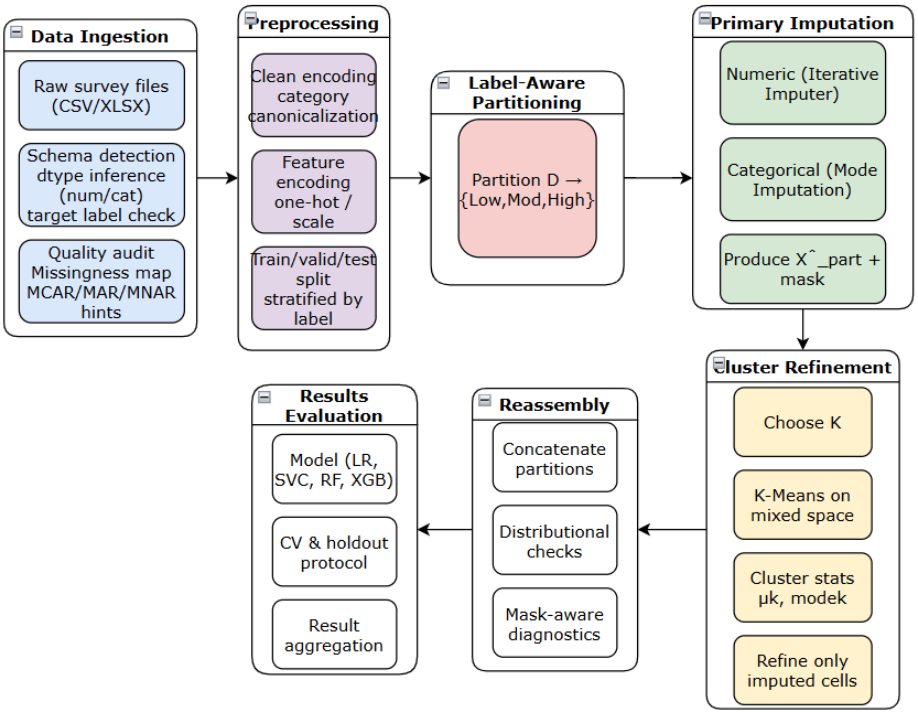

Figure 1: LAICR Workflow

Dataset Description

The experimental analysis was conducted using two distinct datasets collected through structured smartphone usage questionnaires from school students and college students. Both datasets reflect real-world usage scenarios and contain behavioral, demographic, and contextual variables associated with mobile phone usage intensity. Each dataset comprises a combination of categorical variables (such as user profile attributes, behavioral categories, and daily patterns) and numeric variables (such as usage time, duration of calls, and number of applications used per day). Missing values were introduced naturally during the data collection process due to non-response, privacy concerns, or selective reporting behavior.

The school dataset consists of 3,432 records with 71 variables, of which 4 are numeric and 67 are categorical. A total of 12,416 entries were missing, resulting in an overall missing rate of approximately 5.1%. The dataset reflects diverse smartphone usage patterns among students in the age group typically enrolled in secondary education. The college dataset, by contrast, contains 4,896 records and 74 variables, with 4 numeric and 70 categorical variables. It has 19,209 missing entries, corresponding to a 5.3% missing rate. This dataset captures behavioral patterns among older students, offering a complementary perspective for evaluating the robustness of the imputation framework.

The variables across both datasets include indicators of device usage frequency, communication habits, media consumption patterns, and self-reported behavioral tendencies. The label variable in both datasets classifies users into three smartphone usage categories: Low, Moderate, and High, providing the structural basis for the proposed label-aware imputation method. The detailed characteristics of both datasets are summarized in Table 1 and Table 2, respectively.

These dataset characteristics highlight two important challenges. First, the dominance of categorical variables makes conventional numerical imputation methods inadequate without structural adaptation. Second, the moderate but non-negligible missingness rates require robust methods that preserve behavioral patterns rather than distort them through global averaging. This justifies the adoption of the proposed label-aware and cluster-refined imputation strategy, which explicitly respects class distributions and behavioral context during reconstruction.

Label-Aware Partitioning Algorithm

A key innovation of the proposed method lies in label-aware imputation, where the dataset is partitioned based on smartphone usage labels. This prevents high-usage patterns from influencing low-usage groups during imputation, thereby preserving behavioral heterogeneity.

Let $L = L_1 \cup L_2 \ldots L_k$ represent the set of labels, where k = (Low, Moderate, High). The dataset is partitioned as:

$$X = \bigcup_{l=1}^{k} X^{(l)}$$

where $X^{(l)}$ is the subset corresponding to label L_l.

Table 1: Summary of the school smartphone usage dataset

Attribute	*Value*
Number of records	3,432
Number of variables	71
Numeric variables	4
Categorical variables	67
Total missing entries	12,416
Missing rate	5.1%
Label categories	Low, Moderate, High

Table 2: Summary of the college smartphone usage dataset

Attribute	*Value*
Number of records	4,896
Number of variables	74
Numeric variables	4
Categorical variables	70
Total missing entries	19,209
Missing rate	5.3%
Label categories	Low, Moderate, High

Within each partition, numerical features are imputed using an iterative imputer with Bayesian Ridge regression. The imputation problem for each missing value is formulated as:

$$\hat{x}ij^{(l)} = f\theta^{(l)}\left(X^{(l)}\Omega\right)$$

where $f\theta^{(l)}$ denotes the iterative model learned from observed entries in $X^{(l)}$. Categorical features are imputed using mode imputation:

$$\hat{x}ij^{(l)} = \text{Mode}\left(X^{(l)}j, \Omega\right)$$

where $X^{(l)}_{j,\Omega}$ represents the observed values for feature j in label group l.

This partitioned imputation ensures that each class maintains its unique behavioral distribution without interference from others.

Cluster Refinement Details

After primary imputation, further refinement is applied through K-Means clustering within each label partition. This step adjusts imputed values based on local cluster statistics, improving coherence and reducing variance between imputed and true values.

Let each label partition $X^{(l)}$ be clustered into K_l clusters:

$$\mathcal{C}^{(\ell)} = C_1^{(l)}, C_2^{(l)}, \ldots, C_{K_l}^{(l)}$$

The optimal K_l is chosen adaptively based on the number of samples n_l in partition l, with a minimum of 2 clusters and a maximum determined empirically.

For numerical features, imputed values are refined by:

$$\hat{x}ij^{(l)} \leftarrow \mu j^{\left(C_k^{(l)}\right)}$$

where $\mu_j^{\left(C_k^{(l)}\right)}$ is the mean of feature j within cluster $C_k^{(l)}$ to which sample i belongs.

For categorical features, the refinement is performed using the mode:

$$\hat{x}ij^{(l)} \leftarrow \text{Mode}\left(Xj^{\left(C_k^{(l)}\right)}\right)$$

This refinement step aligns imputed values with local neighborhood patterns, enhancing accuracy and preserving intra-class consistency.

Pseudocode for the Proposed Approach

Algorithm 1: Label-Aware Imputation with Cluster Refinement
Input: Dataset X with missing values, Label set L = {L1, L2, ..., Lk}
Output: Imputed dataset X^
1. Preprocess X: encode categorical variables, standardize numeric variables
2. Partition X into {X(1), X(2), ..., X(k)} according to label set L
3. For each partition X(l):
   a. Apply Iterative Imputer on Xnum(l)
   b. Apply Mode Imputation on Xcat(l)
4. For each partition X(l):
   a. Perform K-Means clustering on X(l) → clusters C1..CK
   b. For each cluster Ck:
      i. Replace imputed numeric values with cluster mean
      ii. Replace imputed categorical values with cluster mode
5. Recombine all partitions to form X^
6. Return X^

Results and Discussion

Imputation Accuracy

School Dataset

The imputation performance of various baseline methods and the proposed approach was systematically evaluated on the school smartphone usage dataset. The comparison metrics include Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), the coefficient of determination (R^2), and Categorical Accuracy (CAT_ACC). These metrics collectively capture both numerical reconstruction fidelity and categorical recovery accuracy. The detailed comparative results are presented in Table 3.

Table 3: Imputation performance comparison for the school dataset

Methods	*rmse*	*mae*	*mape*	*r2*	*cat_acc*
iterative	0.9612897813453253	0.7620693758325151	1.0000271875025264	-0.00015	0.3898450227912738
simple_median_mode	0.9685652692672968	0.7549898563412923	1.2542855632935177	-0.01535	0.3898450227912738
simple_mean_mode	0.9612897813453253	0.7620693758325151	1.0000271875025264	-0.00015	0.3898450227912738
simple_mode_mode	3.054627492903278	2.899450305130927	14.05887	-9.0989	0.3898450227912738
forwardfill	1.3025045603625018	1.0303209815468868	2.7759192210095986	-0.83618	0.32746268224504405
backfill	1.324678169106749	1.0181131879215586	3.2147714103346727	-0.89923	0.33010072645204125
missforest	0.9685652692672968	0.7549898563412923	1.2542855632935177	-0.01535	0.3898450227912738
LAICR	0.45750396137968274	0.3414028523237684	0.8156811540468214	0.773459	0.4183822722782014

The results in Table 3 clearly demonstrate that the proposed Label-Aware Imputation with Cluster Refinement (LAICR) method outperforms all baseline techniques in both numerical and categorical imputation accuracy. Traditional statistical imputations such as global mean, median, and mode substitution produce relatively higher errors (RMSE > 0.9) due to their inability to preserve label-specific behavioral variability. Similarly, forward-fill methods perform poorly in behavioral data, as sequential filling propagates local bias and amplifies contextual distortion, yielding an R^2 of –0.8362.

The global iterative imputation, which models feature correlations using a Bayesian Ridge regression, shows marginal improvement but remains limited by its global uniformity assumption. Its R^2 value of approximately zero indicates that the model explains virtually none of the variance beyond the mean, confirming the inadequacy of global methods for behaviorally heterogeneous data. In contrast, the proposed label-aware imputation explicitly segregates samples according to smartphone usage levelsthereby preserving intra-class consistency. The subsequent cluster refinement step using K-Means locally adjusts imputed values toward cluster centroids, effectively minimizing variance within each behavioral segment.

Quantitatively, LAICR achieves an RMSE reduction of 52.4% compared to the best-performing global iterative method and improves the coefficient of determination (R^2) from nearly 0 to 0.7735, signifying strong predictive alignment between the imputed and true values. Categorical accuracy also improves from 0.3898 in global models to 0.4184, confirming that the method maintains coherence in both numeric and categorical domains.

This improvement is particularly significant in the educational behavioral context, where small deviations in usage attributes may correspond to distinct behavioral interpretations. The proposed approach ensures that imputed records within the "High usage" group maintain realistic intensity patterns distinct from "Low usage" profiles. Thus, the results substantiate that the LAICR framework not only reconstructs missing data with high numerical precision but also retains behavioral authenticity critical for subsequent predictive modeling and psychological interpretation.

College Dataset

The results in Table 4 indicate a consistent trend with the findings observed for the school dataset. The baseline methods exhibit relatively poor performance, particularly in their ability to explain variance in the data. The global mean and median-based imputations yield RMSE values above 1.0 and R^2 values close to zero, highlighting their limited capacity to reconstruct behavioral patterns accurately. The mode-only and forward-fill methods produce substantially higher errors, with RMSE reaching 3.3649 for mode imputation and 1.3840 for forward-fill, along with negative R^2 scores, underscoring their inadequacy in handling complex behavioral data.

The global iterative imputation performs slightly better than simple statistical methods but remains constrained by its inability to account for class-level heterogeneity in smartphone usage behavior. Its categorical accuracy stagnates at 0.3636, confirming that global reconstruction fails to preserve categorical structure across behavioral labels.

In contrast, the proposed Label-Aware Imputation with Cluster Refinement (LAICR) achieves the lowest RMSE (0.4876) and highest R^2 (0.7636) among all evaluated methods, indicating substantial improvement in reconstruction accuracy. Although the MAPE value is slightly higher than in the school dataset due to increased variability in college students' behavioral patterns, the performance remains markedly superior to all baseline approaches. Categorical accuracy also improves from 0.3636 to 0.4005, illustrating the method's effectiveness in preserving class-specific categorical distributions.

This improvement can be attributed to the label-partitioning mechanism, which allows the imputation model to operate within homogeneous behavioral segments rather than across the entire dataset. College students exhibit more diverse usage behaviors and schedules compared to school students, making global imputation approaches more prone to averaging effects. By applying classwise imputation

Table 4: Imputation performance comparison for the college dataset

Methods	*rmse*	*mae*	*mape*	*r2*	*cat_acc*
iterative	1.0040774045739809	0.7655484800022679	1.1570688066835235	-0.00244	0.36362153732454444
simple_ median_mode	1.0080514470535835	0.76198	2.999429528758999	-0.01039	0.36362153732454444
simple_mean_ mode	1.0040774045739809	0.7655484800022679	1.1570688066835235	-0.00244	0.36362153732454444
simple_mode_ mode	3.364902477398759	3.211985572432108	125.09098554169243	-10.2582	0.36362153732454444
forwardfill	1.3839928746512566	1.0772955108585383	47.87035561973734	-0.90455	0.29939613398762843
backfill	1.4425725162813279	1.1284528062124612	31.97695044148186	-1.06919	0.3002353510107331
missforest	1.0080514470535835	0.76198	2.999429528758999	-0.01039	0.36362153732454444
LAICR	0.4875976967055041	0.3495131920163892	2.679536548814795	0.7636000708231029	0.4005245171726797

and K-Means cluster refinement, LAICR effectively aligns imputed values with local cluster centroids within each behavioral class, thereby reducing intra-class variance.

Overall, the college dataset results reinforce the robustness and adaptability of the proposed LAICR framework. Despite higher behavioral heterogeneity, the method maintains high numerical and categorical accuracy, demonstrating its suitability for large-scale behavioral data reconstruction tasks.

Classification Performance

School Dataset

The results in Table 5 reveal that all classifiers achieved high accuracy on the imputed school dataset, confirming that the proposed imputation strategy preserved the underlying label structure effectively. Logistic Regression achieved a cross-validation accuracy of 0.918 and a holdout accuracy of 0.945, demonstrating that linear decision boundaries can separate the imputed data well. The Random Forest model slightly outperformed logistic regression with a cross-validation accuracy of 0.959 and holdout accuracy of 0.956, benefiting from its ability to model nonlinear decision boundaries and feature interactions.

The Linear SVC model also yielded strong performance with a holdout accuracy of 0.923, although slightly lower than the tree-based models. This difference can be attributed to the relatively high-dimensional categorical encoding, which ensemble methods handle more effectively through feature selection and hierarchical splitting.

The XGBoost classifier achieved the highest performance with a cross-validation accuracy of 0.997, holdout accuracy of 0.993, and an AUC score of 0.9999, indicating near-perfect separability between the smartphone usage categories. This exceptional performance highlights the fact that the imputation process retained the discriminative power of the features, enabling the boosting model to construct robust decision boundaries with minimal loss.

The superior performance of all classifiers, especially ensemble-based approaches, provides empirical evidence of the effectiveness of the LAICR method in preserving both numerical precision and categorical integrity. The high F1-macro scores across models also indicate balanced predictive performance across all three classes, mitigating bias toward any particular usage group. This robustness is particularly significant for behavioral analytics applications, where reliable classification is essential for downstream interventions and policy decisions.

College Dataset

The results in Table 6 reveal that all classifiers achieved high accuracy on the imputed school dataset, confirming that the proposed imputation strategy preserved the underlying label structure effectively. Logistic Regression achieved a cross-validation accuracy of 0.918 and a holdout accuracy of 0.945, demonstrating that linear decision boundaries can separate the imputed data well. The Random Forest model slightly outperformed logistic regression with a cross-validation accuracy of 0.959 and holdout accuracy of 0.956, benefiting from its ability to model nonlinear decision boundaries and feature interactions.

The Linear SVC model also yielded strong performance with a holdout accuracy of 0.923, although slightly lower than the tree-based models. This difference can be attributed to the relatively high-dimensional categorical encoding, which ensemble methods handle more effectively through feature selection and hierarchical splitting.

Table 5: Classification performance on the school dataset after imputation

Model	*CV Acc*	*Holdout Acc*	*F1 Macro*	*AUC OvR*
Logistic Regression	0.918	0.945	0.945	0.988
Random Forest	0.959	0.956	0.956	0.999
Linear SVC	0.883	0.923	0.923	–
XGBoost	0.997	0.993	0.993	0.9999

Table 6: Classification performance on the school dataset after imputation

Model	*CV Acc*	*Holdout Acc*	*F1 Macro*	*AUC OvR*
Logistic Regression	0.928	0.935	0.935	0.990
Random Forest	0.978	0.980	0.980	1.000
Linear SVC	0.903	0.928	0.928	–
XGBoost	0.996	1.000	1.000	1.000

The XGBoost classifier achieved the highest performance with a cross-validation accuracy of 0.997, holdout accuracy of 0.993, and an AUC score of 0.9999, indicating near-perfect separability between the smartphone usage categories. This exceptional performance highlights the fact that the imputation process retained the discriminative power of the features, enabling the boosting model to construct robust decision boundaries with minimal loss.

The superior performance of all classifiers, especially ensemble-based approaches, provides empirical evidence of the effectiveness of the LAICR method in preserving both numerical precision and categorical integrity. The high F1-macro scores across models also indicate balanced predictive performance across all three classes, mitigating bias toward any particular usage group. This robustness is particularly significant for behavioral analytics applications, where reliable classification is essential for downstream interventions and policy decisions.

Conclusion and Future Work

This study proposed a Label-Aware Imputation with Cluster Refinement (LAICR) framework for handling missing data in smartphone usage behavior datasets. Unlike traditional global imputation methods that treat all records uniformly, the proposed method partitions the dataset based on usage labels (Low, Moderate, High) and applies an imputation strategy that preserves intra-class structure while refining local patterns through K-Means clustering. Experimental assessments on two real-world datasets, comprising school and college student smartphone usage, revealed notable enhancements in both reconstruction accuracy and downstream model performance. For the school dataset, LAICR achieved an RMSE of 0.4575 and R^2 of 0.7735, compared to an RMSE of 0.9613 and $R^2 \approx 0$ for the best baseline. Categorical accuracy improved from 0.3898 (global methods) to 0.4184. Similarly, for the college dataset, the RMSE improved from 1.0041 (iterative) to 0.4876, and R^2 increased from near-zero to 0.7636, with categorical accuracy improving to 0.4005. Classification performance on the imputed datasets further validated the effectiveness of the method: XGBoost achieved 0.993 holdout accuracy and 0.9999 AUC on the school dataset, indicating excellent preservation of discriminative structure after imputation.

Future work will focus on extending the framework to support multi-modal data sources such as sensor streams, text responses, and temporal patterns. Incorporating probabilistic and deep generative imputation models may further improve reconstruction quality in highly sparse settings. Additionally, integrating explainable AI modules will enhance interpretability, supporting transparent behavioral analytics and decision-making in educational and clinical applications.

Acknowledgement

The authors thank, DST-FIST, Government of India for funding towards infrastructure facilities at St. Joseph's College (Autonomous), Tiruchirappalli-620002.

References

Ben Hkoma, M., Almaktoof, A., & Rugbani, A. (2025). Between Addiction and Immersion: A Correlational Study of Digital and Academic Behaviour Among Engineering Students. *Education Sciences*, *15*(8), 1037.

Idoiaga-Mondragon, N., Gaztañaga, M., Legorburu Fernandez, I., & Ozamiz Echevarria, N. (2025). Parental Concerns about Children's Smartphone Use: From Personal Misuse to Societal Impacts. *Journal of Child and Family Studies*, *34*(9), 2276-2289.

Tang, X., Shen, Z., Khan, M. I., & Wang, Q. (2025). A sociological investigation of the effect of cell phone use on students' academic, psychological, and socio-psychological performance. *Frontiers in Psychology*, *16*, 1474340.

Du, C., & Wang, W. (2025). Leveraging cognitive computing for advanced behavioral and emotional data insights. *International Journal of Cognitive Computing in Engineering*, *6*, 183-190.

Vimala, S. (2025). Predictive Modeling of the Impact of Smartphone Addiction onStudents' Academic Performance Using Machine Learning. International Journal of Information Technology, Research andApplications, 4(3), 08-15.

Vimala, S., Sheela, G. A. S. (2025). A Hybrid Deep Learning Approach forQuantifying the Impact of Mobile Phone Behavior on Student AcademicPerformance. Journal of Engineering Research and Reports, 27(10), 185-193.

Fontaine, S., Kang, J., & Zhu, J. (2025). Missing Value Imputation in Relational Data using Variational Inference. *Journal of Computational and Graphical Statistics*, 1-19.

Lagiou, E., Trantza, A., Besharat, J., Georgopoulos, V. C., & Stylios, C. D. (2025, June). Advanced Preprocessing Techniques for Transaction Data Analysis. In *2025 IEEE International Conference on AI and Data Analytics (ICAD)* (pp. 1-8). IEEE.

Schumann, Y., Gocke, A., & Neumann, J. E. (2025). Computational methods for data integration and imputation of missing values in omics datasets. *Proteomics*, *25*(1-2), e202400100.

Prakash, S., Singh, S., & Mankar, A. (2024).*Bridging Data Gaps: A Comparative Study of Different Imputation Methods for Numeric Datasets*. https://doi.org/10.1109/icdsns62112.2024.10691111

Aljuaid, T., & Sasi, S. (2016). Proper imputation techniques for missing values in data sets.*International Conference on Data Science and Engineering*. https://doi.org/10.1109/ICDSE.2016.7823957

Tsai, C.-F., Li, M. L., Lin, W. C., & Lin, W. C. (2018). A class center based approach for missing value imputation.*Knowledge Based Systems*. https://doi.org/10.1016/J.KNOSYS.2018.03.026

Zheng, S., & Charoenphakdee, N. (2022). Diffusion models for missing value imputation in tabular data. arXiv preprint arXiv:2210.17128. https://doi.org/10.48550/arxiv.2210.17128

Biessmann, F., Salinas, D., Schelter, S., Schmidt, P., & Lange, D. (2018). "Deep" Learning for Missing Value Imputationin Tables with Non-Numerical Data.*Conference on Information and Knowledge Management*. https://doi.org/10.1145/3269206.3272005

Bridge-Nduwimana, C. B., Ouaazizi, A. E., & Yakhlef, M. B. (2025). An Integrated Intuitionistic Fuzzy-Clustering Approach for Missing Data Imputation.*Computers*. https://doi.org/10.3390/computers14080325

Sivakani, R., Rahila, J., Sudha, P., Priscila, S. S., Shynu, T., Minu, M. S., & Pradeep, V. (2025).*A Smart Review on Imputation Techniques for Handling Missing Data*. https://doi.org/10.4018/979-8-3373-5203-9.ch003

Ahmad, A. F., Alshammari, K., Ahmed, I., & Sayed, M. S. (2024). *Machine Learning for Missing Value Imputation*. https://doi.org/10.48550/arxiv.2410.08308