Vol. 8, No. 4, August 2026

ISI Journal of Information Systems and Informatics

Published By Asosiasi Doktor Sistem Informasi Indonesia

Explainable XGBoost-Based Prediction of Student Reading Interest Using Digital Learning and Reading Behavior Indicators

Delpiah Wahyuningsih[1], Heki Aprianto[2], Lili Indah Sari[3], Wishnu Aribowo Probonegoro[4], Sri Rahayu[5], Serly Oktarina[6]

[1]Department of Informatics Engineering, [3]Department of Digital Business, [4]Department of Information Systems, ISB Atma Luhur Pangkalpinang, Indonesia

[2]Department of Health Administration, STIKES Budi Mulia Sriwijaya, Palembang, Indonesia

[5]Department of Information Systems, UIN Raden Fatah Palembang, Indonesia

[6]Department of Computer Science, Universitas Sumatera Selatan, Palembang, Indonesia

Received:
February 25, 2026
Revised:
June 3, 2026
Accepted:
August 5, 2026
Published:
August 30, 2026

Corresponding Author:
Author Name*:
Author Name
Email*:
Author Email

DOI:
10.63158/journalisi.v8i4.1698

© 2026 Journal of Information Systems and Informatics. This open access article is distributed under a (CC-BY License)

Abstract. Reading interest is an important indicator of academic engagement and lifelong learning, yet its relationship with digital learning behavior remains underexplored in higher education. This study proposes an explainable machine learning framework to predict student reading interest using digital learning behavior, reading habits, and demographic characteristics. Data were collected through self-reported questionnaires from 300 Indonesian university students. Reading interest was formulated as a binary target variable, and three feature scenarios were evaluated: digital behavior only (S1), reading behavior only (S2), and an integrated model combining all features (S3). The dataset was divided using an 80:20 stratified train–test split, with SMOTE applied only to the training data to address class imbalance while preventing information leakage. S3 achieved the highest ROC-AUC (91.92%) and cross-validation F1-score (86.62% ± 3.34%), while S1 obtained the same test accuracy (84.29%) and a slightly higher test F1-score (85.33%). SHAP analysis provided interpretability by identifying daily reading duration and monthly book count as key predictors. These findings indicate that multidimensional behavioral indicators can support early student-engagement screening systems. However, results from a single institutional context require cautious interpretation.

Keywords: Reading Interest; Educational Data Mining; XGBoost; Explainable AI; SHAP; Digital Learning Behavior

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

1. INTRODUCTION

Reading interest is widely recognized as a fundamental driver of academic success and intellectual development among students. In the contemporary digital era, the landscape of reading has undergone substantial transformation, with students increasingly engaging with e-books, online articles, and digital platforms as primary sources of information. Despite this shift, the relationship between students' digital learning behaviors and their reading interest levels remains underexplored in the educational data mining and learning analytics literature [1] [2] [3].

Reading interest is commonly understood as an individual's tendency, willingness, and motivation to voluntarily engage in reading-related activities, often associated with intrinsic motivation and sustained literacy engagement. In contrast, reading habit reflects repeated behavioral patterns related to reading frequency and duration, whereas reading engagement refers to broader cognitive, emotional, and behavioral involvement in literacy activities. Within digitally mediated educational settings, digital learning behavior encompasses students' interactions with digital resources and platforms including e-books, online articles, learning-management systems, and educational digital content that may indirectly shape literacy practices and reading preferences. From the perspective of reading-motivation theory, students who possess stronger intrinsic motivation are more likely to voluntarily engage in sustained reading activities, thereby strengthening long-term literacy engagement.

In higher education, reading interest represents an important predictive target because it may function as an early indicator of academic engagement, self-regulated learning, and sustained literacy development. Rather than serving solely as a descriptive construct, understanding reading interest may support preliminary identification of students who potentially demonstrate lower engagement with reading-related activities, thereby informing educator awareness in digitally intensive learning environments.

Artificial intelligence and machine learning have demonstrated considerable promise in supporting educational decision-making by enabling predictive models that can identify students at risk and inform targeted interventions [4] [5] . Numerous studies have applied machine learning to predict academic outcomes, dropout risk, and student performance

by leveraging demographic, academic, and behavioral attributes [6] [7] [8] [9]. However, the specific prediction of reading interest based on digital behavior patterns represents a gap in the existing body of work.

Several digital-learning indicators were considered in this study based on their potential association with student literacy engagement. For example, social-media usage duration may provide indirect behavioral signals because students use social-media platforms for both educational and non-educational purposes. Consequently, its relationship with reading interest may depend on the type of content students consume rather than the amount of time spent on social media alone. Likewise, LMS usage frequency, online learning duration, online article access, and internet quality may influence students' opportunities to access digital learning resources and participate in academic reading activities, whereas digital entertainment duration may reflect competing patterns of time allocation that indirectly influence reading behavior. Accordingly, this study distinguishes educational digital activities (e.g., LMS usage, online learning, e-books, and online articles) from general digital exposure (e.g., social media and digital entertainment), recognizing that these categories may contribute differently to reading-interest prediction.

In addition to digital-learning indicators, this study includes daily reading duration and monthly book count as behavioral predictors rather than direct measures of reading interest. These variables represent observable reading habits that may reflect students' engagement with reading activities but do not define reading interest itself. Although students with stronger reading interest are generally expected to spend more time reading and complete more books, these behaviors are modeled as observable manifestations associated with reading interest rather than variables used to define the target construct. This distinction allows the model to examine how observable reading behavior contributes to predicting self-reported reading interest while acknowledging the potential conceptual overlap between the predictors and the target variable.

A critical limitation of many machine learning applications in education is their reliance on black-box models, which sacrifice transparency for predictive accuracy [10] [11]. Explainable Artificial Intelligence (XAI) has emerged as a methodological paradigm addressing this gap by providing interpretable explanations for model predictions.

Techniques such as SHAP (SHapley Additive exPlanations) and LIME enable both global feature attribution and instance-level explanations [12] [13] [14].

This study addresses the aforementioned gaps by developing an interpretable machine learning framework for predicting student reading interest based on digital learning and reading behavior indicators. Three distinct feature scenarios are evaluated: digital behavior only (S1), reading behavior only (S2), and a combined scenario integrating digital behavior, reading behavior, and demographic features (S3). XGBoost is employed as the predictive model, and SHAP is integrated to ensure model transparency. The proposed framework is intended to support preliminary student-engagement screening and early identification of students who may exhibit lower reading engagement and therefore benefit from additional educator attention. Accordingly, explainable AI is employed to improve predictive transparency and support informed educational decision-making rather than automated intervention.

To address the identified research gaps, this study investigates the predictive relationship between digital learning behavior, reading-habit indicators, and student reading interest using an explainable machine learning approach. Specifically, this study addresses the following research questions:

1) RQ1: Which feature group digital learning behavior, reading-habit indicators, or their integration with demographic attributes provides the strongest predictive performance for student reading interest?
2) RQ2: Which behavioral and demographic indicators contribute most substantially to student reading interest prediction?
3) RQ3: How can explainable machine learning improve interpretability and transparency in student reading-interest prediction?

The contributions of this study are structured into three dimensions. Conceptually, this study extends educational data mining by integrating concepts from literacy engagement theory with digital learning behavior and explainable machine learning into a unified predictive framework for student reading-interest prediction. Methodologically, this study combines scenario-based feature evaluation with explainable XGBoost modeling to compare the predictive contribution of digital learning behavior, reading behavior, and demographic characteristics while maintaining model transparency through SHAP

explanations. Practically, the proposed framework provides an interpretable decision-support approach for preliminary student-engagement screening while preserving educator oversight in educational decision-making.

2. METHODS

Figure 1 illustrates the overall workflow of the proposed study, beginning with questionnaire-based data collection, followed by data preprocessing, model development, model evaluation, SHAP-based explainability analysis, and interpretation of the findings.

Figure 1. Research workflow of the proposed explainable XGBoost framework for predicting student reading interest

Figure 1 illustrates the overall methodology employed in this study. The process began with a cross-sectional questionnaire survey administered to undergraduate students. The collected responses were inspected, cleaned, encoded, and transformed into machine-learning features. The dataset was subsequently divided into stratified training and testing subsets. To address class imbalance, SMOTE was applied exclusively to the training subset, while the testing subset remained unchanged to prevent information leakage. Three XGBoost scenarios were then developed and evaluated using Accuracy, Precision,

Recall, F1-score, ROC-AUC, and five-fold cross-validation. Finally, SHAP was employed to explain model predictions through global feature importance, beeswarm plots, dependence plots, and force plots, followed by interpretation of the findings, educational implications, study limitations, and future research directions.

2.1. Participants and Data Collection

The dataset consists of 300 university student records collected from higher education institutions in Indonesia through a structured questionnaire survey. Respondents were selected using convenience sampling based on the following inclusion criteria: (1) active university students, (2) regular engagement with digital learning activities, and (3) willingness to voluntarily participate in the study. Data collection was conducted during the 2025 academic period.

The questionnaire captured three categories of variables: digital learning behavior, reading-habit indicators, and demographic characteristics. All behavioral indicators were obtained through self-reported responses, rather than objective learning-management-system logs or automatically captured digital traces. To ensure participant privacy, all responses were anonymized before analysis, and no personally identifiable information was retained. Participation was voluntary and conducted under informed consent for research purposes.

2.2. Questionnaire Design

The questionnaire was designed as a structured data collection instrument to obtain behavioral, demographic, and reading-interest information for machine-learning analysis. Rather than serving as a newly developed psychometric measurement scale, the questionnaire was intended to collect observable behavioral indicators and contextual characteristics associated with students' reading activities. Accordingly, the collected responses were used as predictive input variables rather than latent psychological constructs.

The questionnaire consisted of four sections: (1) demographic characteristics, (2) digital learning behavior, (3) reading behavior, and (4) reading-interest assessment. Demographic information included gender, semester, and study program. Digital learning behavior covered variables such as social media usage duration, digital entertainment duration, e-

book usage frequency, online article reading frequency, online learning duration, smartphone-assisted learning frequency, learning management system (LMS) usage frequency, and internet quality. Reading behavior included daily reading duration, monthly book count, reading location, reading media, preferred book genre, and reading motivation.

Most behavioral variables were measured using ordinal response categories representing frequency or duration levels, whereas demographic variables were represented as binary or nominal categories where appropriate. Reading interest was assessed using a single self-reported item, "How would you rate your interest in reading books?", with four response options: Dislike, Neutral, Like, and Strongly Like. As described in Section 2.3, these responses were subsequently transformed into a binary target variable for the prediction task. A complete description of all predictor variables, response scales, data types, and encoding methods is provided in Table 1 (Feature Dictionary).

2.3. Reading Interest Target Construction

The target variable representing student reading interest was constructed from an independent self-assessment questionnaire item: *"How would you rate your interest in reading books?"* Responses were transformed into a binary classification label. Students answering "Like" or "Strongly Like" were categorized as High Reading Interest, whereas those answering "Neutral" or "Dislike" were categorized as Low Reading Interest. To minimize potential target leakage, the reading-interest label was constructed independently from predictor preprocessing. However, because several reading-habit indicators may conceptually overlap with reading-interest behavior, this issue is acknowledged as a study limitation and discussed further in the Discussion section.

2.4. Predictor Variables

The predictor variables used in this study were grouped into three categories: digital learning behavior, reading-habit indicators, and demographic characteristics. Digital learning behavior variables included social-media usage duration, digital entertainment duration, e-book frequency, online article frequency, online learning duration, smartphone learning frequency, learning-management-system (LMS) usage frequency, and internet quality. Reading-habit indicators included daily reading duration, monthly book count, reading location, reading media, book genre, and reading motivation, whereas

demographic variables consisted of gender, semester, and study program. Most behavioral variables were measured using ordinal response categories, while contextual and demographic variables were treated as nominal or binary attributes where appropriate. Prior to model training, categorical responses were transformed through label encoding to ensure compatibility with machine-learning processing and XGBoost model requirements.

Table 1 presents the complete feature dictionary used in this study, including the description, response scale, data type, and encoding method for each predictor variable. Binary variables, such as gender, were encoded using 0–1 representation. Ordinal variables were encoded according to their natural ordering of response categories, whereas nominal variables were transformed using label encoding. Although one-hot encoding is commonly recommended for nominal variables, label encoding was adopted because the XGBoost algorithm is tree-based and relatively insensitive to monotonic integer representations of categorical features while maintaining a compact feature space for the relatively small dataset.

Table 1. Feature Dictionary

Variable	Description	Response Scale	Data Type	Encoding
Gender	Student gender	Male, Female	Binary	0-1
Semester	Current semester	Semester 1–8	Ordinal	Integer
Study Program	Student study program	SI, TI, BD	Nominal	Label Encoding
Social Media Duration	Daily social media usage	<1 h, 1–2 h, 2–4 h, >4 h	Ordinal	Label Encoding
Digital Entertainment Duration	Daily digital entertainment usage	<1 h, 1–2 h, 2–4 h, >4 h	Ordinal	Label Encoding
E-book Frequency	Frequency of reading e-books	Never, Rarely, Sometimes, Often, Always	Ordinal	Label Encoding
Online Article Frequency	Frequency of reading online articles	Never, Rarely, Sometimes, Often, Always	Ordinal	Label Encoding

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

Variable	Description	Response Scale	Data Type	Encoding
Online Learning Duration	Daily online learning duration	<1 h, 1–2 h, 2–4 h, >4 h	Ordinal	Label Encoding
Smartphone Learning Frequency	Frequency of smartphone-assisted learning	Never, Rarely, Sometimes, Often, Always	Ordinal	Label Encoding
LMS Usage Frequency	Frequency of LMS usage	Never, Rarely, Sometimes, Often, Always	Ordinal	Label Encoding
Internet Quality	Perceived internet quality	Poor, Fair, Good, Excellent	Ordinal	Label Encoding
Daily Reading Duration	Daily reading duration	<30 min, 30–60 min, 1–2 h, >2 h	Ordinal	Label Encoding
Monthly Book Count	Number of books read per month	0, 1–2, 3–5, >5	Ordinal	Label Encoding
Reading Location	Preferred reading location	Home, Library, Campus, Others	Nominal	Label Encoding
Reading Media	Preferred reading media	Printed, Digital, Both	Nominal	Label Encoding
Book Genre	Preferred book genre	Academic, Fiction, Non-fiction, Others	Nominal	Label Encoding
Reading Motivation	Self-reported reading motivation	Very Low, Low, Moderate, High, Very High	Ordinal	Label Encoding

2.5. Preprocessing and Class Balancing

The collected dataset was first inspected for duplicate records, missing values, and inconsistent responses prior to model development. Since the questionnaire was administered using mandatory response fields, no missing values were observed in the final dataset. Categorical predictor variables were transformed into numerical representations using label encoding as described in Section 2.4.

The dataset consisted of 300 student records, comprising 175 samples of Low Reading Interest and 125 samples of High Reading Interest, indicating a moderate class imbalance. To preserve the original class distribution, the dataset was partitioned into training and

testing subsets using an 80:20 stratified train-test split with a fixed random seed of 42. Consequently, the training set contained 140 Low Reading Interest and 100 High Reading Interest instances, while the testing set contained 35 Low Reading Interest and 25 High Reading Interest instances.

To mitigate the effect of class imbalance during model training, the Synthetic Minority Oversampling Technique (SMOTE) was applied exclusively to the training subset [15], generating synthetic samples for the minority class until both classes contained an equal number of observations. After oversampling, the training dataset consisted of 140 Low Reading Interest and 140 High Reading Interest samples, whereas the testing dataset remained unchanged to ensure an unbiased performance evaluation.

Although SMOTENC is commonly recommended for datasets containing mixed categorical and continuous variables, all predictor variables in this study were transformed into numerical encoded representations prior to model training. Therefore, standard SMOTE was employed to balance the training data while maintaining a consistent preprocessing workflow. Future studies may investigate the comparative performance of SMOTENC for mixed-feature educational datasets. The overall preprocessing workflow consisted of categorical encoding, stratified train-test splitting, class balancing using SMOTE on the training subset, and subsequent model training using the balanced training data.

Table 2.

Dataset	Low Reading Interest	High Reading Interest
Original Dataset	175	125
Training Set	140	100
Training Set (After SMOTE)	140	140
Testing Set	35	25

2.6. Model and Evaluation

XGBoost (Extreme Gradient Boosting) was selected as the primary classification algorithm due to its strong performance in handling tabular datasets, nonlinear feature interactions, and imbalanced classification problems, while maintaining computational efficiency and compatibility with explainable machine-learning frameworks [16]. The

dataset was split into training (80%) and test (20%) subsets using stratified sampling. Model performance was evaluated using Accuracy, Precision, Recall, F1-Score, ROC-AUC, and 5-fold cross-validation F1-Score to assess both predictive performance and model stability.

The XGBoost classifier was configured using the following parameters: n_estimators = 150, max_depth = 4, learning_rate = 0.05, subsample = 0.8, colsample_bytree = 0.8, eval_metric = "logloss", and random_state = 42. These hyperparameters were manually specified based on preliminary experimentation and recommendations from previous studies. No exhaustive hyperparameter optimization, such as Grid Search, Random Search, or Bayesian Optimization, was performed in this study. Following model training, model interpretability was achieved using SHapley Additive exPlanations (SHAP), which quantify the contribution of each predictor variable to individual predictions and provide both global and local explanations of the XGBoost model.

2.7. SHAP Explainability

SHAP (SHapley Additive exPlanations) [14] was integrated to provide multi-level interpretability for the XGBoost model. Explanations were generated using the TreeExplainer implementation to ensure compatibility with the XGBoost architecture. SHAP values were interpreted as predictive feature contributions within the model output space, rather than causal effects or intervention estimates. Global explanations were generated through SHAP beeswarm plots to identify influential predictors across all observations. Local interpretability was examined using SHAP force plots for representative prediction samples, while SHAP dependence plots were constructed to visualize predictive tendencies among the most influential variables.

2.8. Experimental Environment

The experiments were conducted using the Kaggle Notebook cloud computing environment with Python 3.12.12. Machine-learning implementation was performed using XGBoost 3.2.0**,** Scikit-learn 1.6.1**,** SHAP 0.50.0**,** Pandas 2.3.3**,** NumPy 2.0.2, and imbalanced-learn 0.14.1. To improve reproducibility, all stochastic processes, including dataset partitioning, SMOTE oversampling, and model initialization, were executed using a fixed random seed of 42. Table X summarizes the experimental environment used in this study.

Table 3. Experimental environment

Component	Version
Platform	Kaggle Notebook
Python	3.12.12
XGBoost	3.2.0
Scikit-learn	1.6.1
SHAP	0.50.0
Pandas	2.3.3
NumPy	2.0.2
imbalanced-learn	0.14.1
Random Seed	42

3. RESULTS AND DISCUSSION

3.1. Dataset Distribution and Experimental Setup

Following preprocessing, the final dataset consisted of 300 valid student records. The target-class distribution prior to splitting comprised 175 students categorized as Low Reading Interest and 125 students categorized as High Reading Interest. Using a stratified 80:20 train–test split, the dataset was partitioned into 240 training records and 60 testing records, preserving class proportions across both subsets. To address class imbalance, SMOTE was applied exclusively to the training data, increasing the minority-class representation to achieve a balanced class distribution prior to model training. No resampling was performed on the testing data to avoid information leakage and ensure unbiased model evaluation.

3.2. Model Performance Comparison

The predictive performance of the three experimental scenarios is summarized in Table 1. Among the evaluated scenarios, Scenario 3 (S3) demonstrated the highest ROC-AUC (91.92%) and strongest cross-validation F1-score stability (86.62% ± 3.34%), indicating improved discrimination capability and model robustness across data partitions. However, Scenario 1 (S1) achieved the same test accuracy (84.29%) and a slightly higher test F1-score (85.33%), suggesting that digital learning behavior alone already provides strong predictive signals of student reading interest. Therefore, S3 is interpreted not as an

unequivocally superior model, but rather as the most balanced scenario in terms of predictive robustness, multidimensional feature integration, and cross-validation stability.

Table 4. Model Performance Comparison Across Three Scenarios

Metric	S1: Digital	S2: Reading Behavior	S3: All Features
Accuracy	84.29%	78.57%	**84.29%**
F1-Score	**85.33%**	78.87%	84.06%
ROC-AUC	88.90%	85.84%	**91.92%**
CV F1 (5-fold)	82.56% ± 3.79%	78.49% ± 1.67%	**86.62% ± 3.34%**

Interestingly, Scenario 1 (Digital Only) achieved stronger predictive performance than Scenario 2 (Reading Behavior Only), with an accuracy improvement of 5.72 percentage points and a higher ROC-AUC (88.90% vs. 85.84%). This result suggests that students' digital learning activities contain stronger predictive signals of reading interest than conventional reading behavior indicators alone. Such findings may reflect the ongoing transformation of reading culture in higher education, where students increasingly consume academic and informational content through digital ecosystems, including online articles, e-books, learning management systems, and educational social media platforms.

Furthermore, the relatively low standard deviation observed in the 5-fold cross-validation results indicates that the proposed models maintain stable predictive performance across different data partitions, thereby reducing the likelihood of overfitting. In particular, S3 exhibited superior consistency compared with S1 and S2, further reinforcing the robustness of combining digital, reading, and demographic variables for student reading interest prediction.

The classification report for S3 also indicates balanced predictive sensitivity across both target classes, with recall values of 0.86 for low reading interest and 0.83 for high reading interest. This balanced performance demonstrates that the model does not disproportionately favor one class over another, thereby improving reliability in identifying both highly engaged readers and students with potentially low reading engagement.

3.3. SHAP Feature Importance Analysis

Global SHAP analysis for Scenario 3 identified **daily reading duration** (*Durasi Baca/hari*) and **monthly book count** (*Jumlah Buku/bulan*) as the two most influential predictors of student reading interest. The dominance of these variables suggests that reading engagement remains fundamentally associated with sustained reading behavior and reading frequency.

Table 5. Top 10 Most Influential Features Based on SHAP Values (Scenario 3)

Rank	Feature	Mean |SHAP|	Feature Group
1	Daily Reading Duration	0.8441	Reading
2	Monthly Book Count	0.6136	Reading
3	Social Media Duration	0.4237	Digital
4	E-book Frequency	0.3655	Digital
5	Online Article Frequency	0.3528	Digital
6	Study Program	0.3395	Demographic
7	Reading Motivation	0.2811	Reading
8	Digital Entertainment Duration	0.2756	Digital
9	Semester	0.2492	Demographic
10	Gender	0.2433	Demographic

Specifically, daily reading duration may reflect persistence and cognitive engagement in literacy-related activities, while monthly book count may indicate habitual reading commitment and voluntary learning behavior. Beyond traditional reading indicators, several digital behavior variables emerged among the top predictors, including social media usage duration, e-book frequency, and online article reading frequency. The prominence of these variables highlights the growing role of digital ecosystems in shaping reading habits among university students. Rather than replacing conventional reading practices, digital engagement appears to complement reading behavior [17], particularly when students interact with educational or informational content through digital media.

Interestingly, demographic attributes such as program of study, semester, and gender also contributed meaningfully to prediction performance, suggesting that reading interest may be influenced by contextual educational characteristics. This finding reinforces the multidimensional nature of student reading engagement, indicating that reading interest emerges through the interaction of behavioral and demographic factors rather than isolated variables alone.

3.4. SHAP Beeswarm Plot Analysis

Figure 2 presents the SHAP beeswarm plot for Scenario 3, illustrating the global distribution of feature contributions toward student reading interest prediction. The beeswarm visualization provides insight into both the magnitude and direction of feature influence across all observations, where positive SHAP values contribute toward high reading interest, whereas negative SHAP values indicate a tendency toward low reading interest.

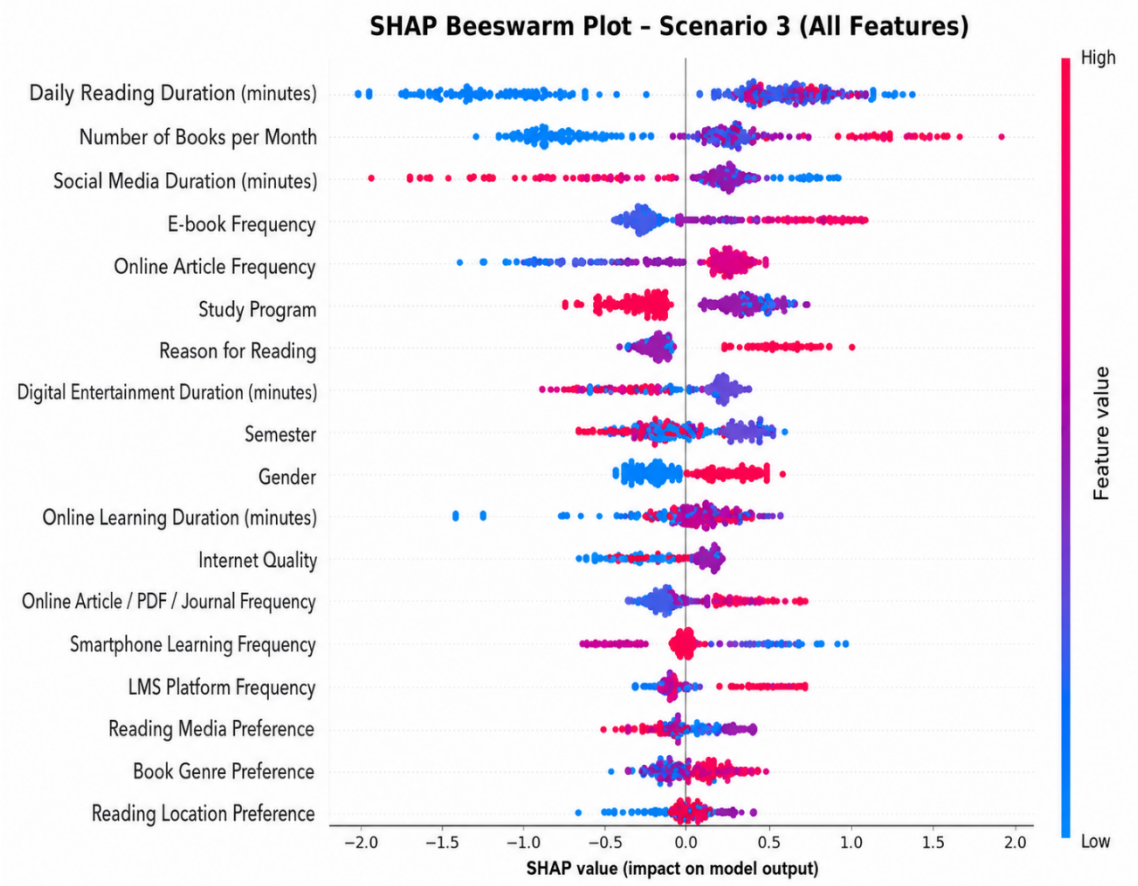

Figure 2. SHAP Beeswarm Plot for Scenario 3 (All Features)

The plot reveals that daily reading duration (*Durasi Baca/hari*) exhibits the strongest and most consistent contribution to prediction outcomes. Higher feature values are predominantly associated with positive SHAP contributions, while lower values strongly shift predictions toward lower reading interest. This clear monotonic pattern suggests that sustained daily reading engagement functions as the most influential behavioral proxy for student reading interest.

A similarly strong trend is observed for monthly book count (*Jumlah Buku/bulan*), where students reporting higher reading frequency consistently demonstrate positive SHAP values. The distribution indicates that students reading more books per month tend to be strongly associated with high reading interest predictions, reinforcing the importance of reading quantity as an indicator of literacy engagement.

Among digital behavior variables, social media usage duration (*Durasi Media Sosial*) demonstrates a more heterogeneous SHAP distribution, where high feature values contribute to both positive and negative prediction outcomes. This pattern suggests that social media engagement does not exert a uniformly detrimental effect on reading interest. Instead, its contribution appears context-dependent, potentially influenced by how students utilize digital platforms. Students engaging with informational or educational content through social media may potentially exhibit different reading-related behaviors. however, because the purpose and context of platform usage were not directly measured, this interpretation should be approached cautiously.

Other digital behavior indicators, including e-book frequency (*Frek. E-book*) and online article frequency (*Frek. Artikel Online*), demonstrate predominantly positive SHAP contributions at higher feature values. These findings suggest that digital reading practices complement traditional reading behavior, reflecting the transition toward digitally mediated literacy in higher education environments. Interestingly, contextual demographic variables such as program of study, semester, and gender also contribute to prediction performance, although with lower magnitude compared to behavioral indicators. Their presence among influential predictors suggests that reading interest may vary according to educational context and student background characteristics. Moreover, lower-ranked features including reading location, book genre, and reading media exhibit SHAP values clustered near zero, indicating comparatively smaller

contributions to model prediction. This finding implies that while contextual preferences may influence reading habits, sustained reading behavior and digital engagement remain substantially stronger predictors of student reading interest.

Moreover, lower-ranked features including reading location, book genre, and reading media exhibit SHAP values clustered near zero, indicating comparatively smaller contributions to model prediction. This finding implies that while contextual preferences may influence reading habits, sustained reading behavior and digital engagement remain substantially stronger predictors of student reading interest. Overall, the SHAP beeswarm visualization highlights the multidimensional nature of reading interest prediction, emphasizing that student literacy engagement emerges from interactions between conventional reading habits, digital learning behavior, and contextual educational characteristics.

3.5. SHAP Force Plot Analysis

To provide instance-level interpretability, SHAP force plots were generated for representative prediction samples. Representative samples were selected to illustrate diverse prediction outcomes, including correctly predicted cases of both high and low reading interest, thereby enabling clearer interpretation of how multiple variables jointly contribute to model predictions. These plots decompose individual predictions into additive feature contributions relative to the baseline prediction, thereby enabling transparent examination of how specific variables influence predicted reading interest.

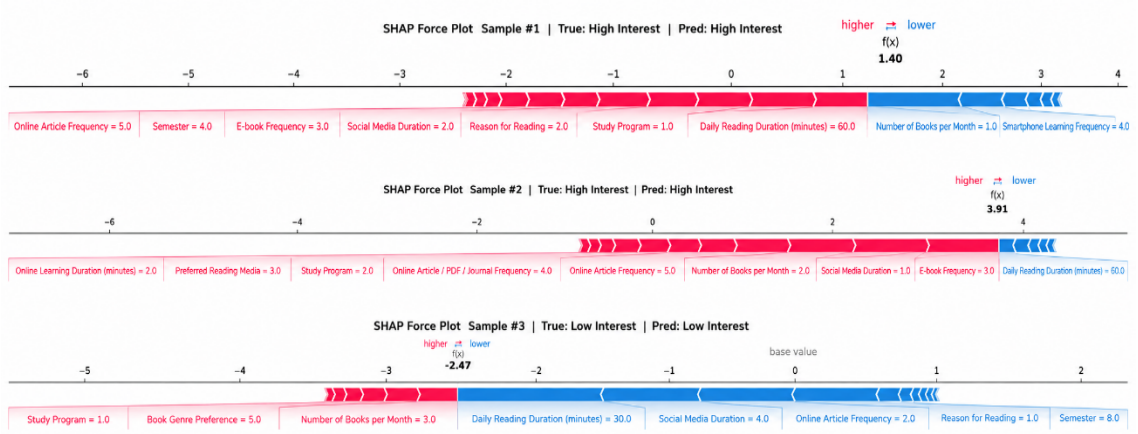

Figure 3. SHAP Force Plots for Representative Prediction Samples

For Sample #1, classified as high reading interest, positive contributions from online article frequency, semester, e-book frequency, and daily reading duration outweighed negative contributions from monthly book count and smartphone learning frequency, ultimately producing a positive prediction score. This result illustrates how digital engagement variables can compensate for relatively weaker traditional reading indicators. Sample #2 demonstrated an even stronger positive prediction, driven by a combination of high online learning duration, e-book frequency, reading duration, and monthly book count, suggesting complementary interactions between digital learning activities and conventional reading behavior.

Conversely, Sample #3, correctly classified as low reading interest, was characterized by low daily reading duration, reduced online article frequency, and limited motivational indicators, collectively contributing toward a strongly negative prediction. This finding reinforces the importance of sustained reading behavior in differentiating high-interest and low-interest students. More importantly, these local explanations demonstrate the practical utility of explainable AI in educational settings. Rather than providing only generalized population-level predictions, the model enables educators to better understand student-specific behavioral patterns associated with reading interest. Accordingly, the framework is intended to support preliminary student-engagement screening and educator interpretation, rather than direct or automated intervention decisions.

3.6. SHAP Dependence Plot Analysis

Figures 4 and 5 present SHAP dependence plots for the two most influential predictors: daily reading duration and monthly book count. Figure 4 illustrates a threshold-like predictive pattern between daily reading duration and model output. Students reporting lower daily reading duration generally exhibit negative SHAP contributions, indicating a stronger association with predictions of lower reading interest. As daily reading duration increases, SHAP contributions tend to shift toward positive values, suggesting a greater association with predictions of higher reading interest. Furthermore, the interaction pattern with monthly book count indicates that students who report both longer reading duration and higher reading frequency generally exhibit stronger positive SHAP contributions. This observation suggests a complementary predictive relationship between reading duration and reading quantity within the analyzed dataset.

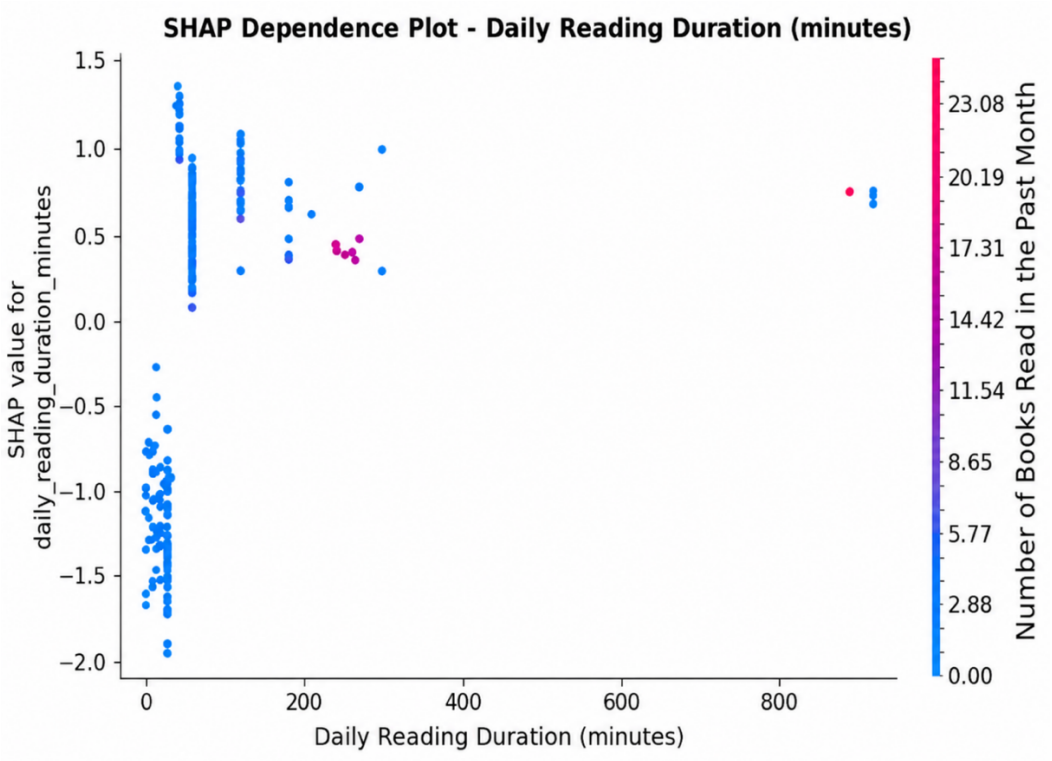

Figure 4. SHAP Dependence Plot – Daily Reading Duration (minutes)

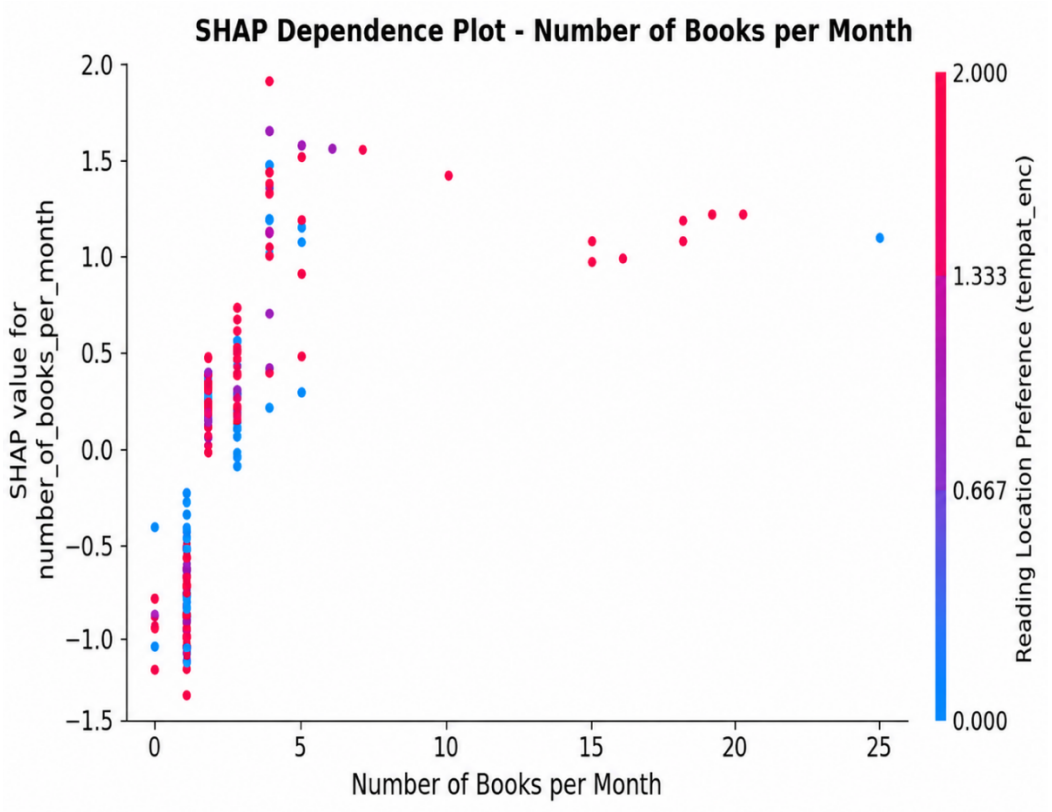

Figure 5. SHAP Dependence Plot – Monthly Book Count

Similarly, Figure 5 demonstrates a positive predictive relationship between monthly book count and model prediction. Students reporting lower monthly reading frequency tend to exhibit more negative SHAP contributions, whereas students with higher reported book consumption generally show more positive SHAP values. These findings suggest that reading quantity may serve as a meaningful predictive indicator associated with student reading interest, particularly when interpreted alongside daily reading behavior.

It is important to note that SHAP values represent predictive contributions rather than causal relationships. Accordingly, variables such as daily reading duration and monthly book count should be interpreted as strong predictive indicators associated with reading interest, rather than direct determinants that necessarily improve reading-interest outcomes. Likewise, the threshold-like patterns observed in dependence plots are interpreted as predictive tendencies within the analyzed dataset, rather than validated behavioral recommendations for educational intervention.

3.7. Discussion

The findings of this study provide several important contributions to educational data mining and explainable machine learning research. First, the superior performance of Scenario 3 demonstrates that student reading interest is more effectively predicted when digital learning behavior, reading behavior, and demographic characteristics are considered together. Rather than suggesting that any single factor independently determines reading interest, these findings indicate that reading engagement is a multidimensional phenomenon emerging from the interaction of multiple behavioral and contextual characteristics. This observation is consistent with previous educational data mining studies reporting that heterogeneous feature integration often improves predictive robustness.

Second, the observation that the Digital Only scenario outperformed the Reading Behavior Only scenario suggests that digital learning activities contain meaningful predictive information associated with student reading interest. However, this finding should not be interpreted as evidence that digital behavior is inherently more important than conventional reading behavior. Instead, it may reflect the increasing integration of reading activities into digital environments, where students frequently access academic

materials through e-books, online articles, learning management systems, and other educational platforms.

Importantly, the SHAP analysis identified daily reading duration and monthly book count as the strongest predictive contributors. Nevertheless, these variables should be interpreted with caution because they may conceptually overlap with the target variable representing reading interest. Students who report greater interest in reading are naturally more likely to spend longer periods reading and to complete more books each month. Consequently, these variables may represent behavioral manifestations associated with reading interest rather than completely independent predictors. Their high SHAP importance therefore reflects predictive association within the learned model rather than evidence of causal influence.

Similarly, the observed SHAP dependence patterns should not be interpreted as behavioral thresholds or intervention targets. The transition from negative to positive SHAP values represents predictive tendencies learned from the analyzed dataset rather than validated behavioral cut-off points. Accordingly, increasing reading duration or monthly book count cannot be assumed to directly improve student reading interest without additional longitudinal or experimental evidence.

The contribution of social media usage also requires careful interpretation. Although social media duration appeared among the influential predictors, the questionnaire did not distinguish whether students primarily consumed educational, informational, or entertainment-oriented content. Consequently, the results should not be interpreted as evidence that increasing social media use positively influences reading interest. Instead, the findings merely indicate that social media behavior contains predictive information under the specific conditions represented in the collected dataset.

The integration of SHAP substantially enhances model transparency by providing both global and instance-level explanations. Nevertheless, these explanations should not be interpreted as behavioral prescriptions for educational intervention. SHAP values quantify the contribution of individual variables to model predictions rather than identifying factors that should necessarily be modified to improve student outcomes.

Therefore, explainable AI should be regarded as a decision-support mechanism that complements, rather than replaces, educator expertise and professional judgment.

The inclusion of demographic variables such as gender, semester, and study program also raises important ethical considerations. These variables were incorporated solely to improve predictive representation and should not be used as independent criteria for labeling, ranking, or making educational decisions about individual students. Any practical implementation should ensure that model outputs remain subject to educator review, contextual interpretation, and institutional oversight to minimize the risk of unintended bias or stigmatization. Likewise, appropriate privacy protection, informed consent, and secure management of student behavioral data should remain essential requirements for any future deployment of predictive educational systems.

Finally, although XGBoost demonstrated strong predictive performance in this study, future research should compare the proposed framework with simpler and inherently interpretable models, such as Logistic Regression and Decision Trees, to further evaluate the trade-off between predictive performance and interpretability. Such comparisons would help determine whether similar predictive capability can be achieved using less complex models while maintaining transparency and ease of practical adoption in educational settings.

3.8. Limitations

Several limitations should be acknowledged. First, the dataset relies on self-reported questionnaire responses, which may introduce response bias and measurement subjectivity. Furthermore, the questionnaire was designed as a structured data collection instrument for predictive modeling and was not subjected to formal psychometric validation, such as construct validity or internal consistency reliability testing. Therefore, the collected responses should be interpreted as behavioral indicators rather than standardized psychological measurements. Future studies should validate the questionnaire using established psychometric procedures before broader application.

Second, although the reading-interest target was constructed independently from predictor preprocessing, several behavioral indicators, particularly reading duration and monthly book count may conceptually overlap with the reading-interest construct,

potentially influencing predictive interpretation. Third, the study was conducted within a limited higher-education context, which may restrict generalizability to other institutions or educational settings. Fourth, the study evaluates predictive performance rather than intervention effectiveness, and therefore the findings should not be interpreted as evidence that modifying specific behaviors will directly improve student reading interest. Finally, objective behavioral signals derived from learning-management systems or digital logs were not available and should be explored in future work. Future studies should additionally compare the proposed framework with simpler interpretable models to further evaluate trade-offs between predictive performance and explainability.

4. CONCLUSION

This study proposed an explainable machine learning framework for predicting student reading interest using digital learning behavior, reading-habit indicators, and demographic characteristics within a higher-education context. Across the evaluated experimental scenarios, Scenario 3 (S3) achieved the highest ROC-AUC (91.92%) and a cross-validation F1-score of 86.62% ± 3.34%, whereas Scenario 1 (S1) achieved the same test accuracy (84.29%) and a slightly higher test F1-score (85.33%), indicating that digital learning behavior alone provides meaningful predictive signals of student reading interest. SHAP-based explainability further enhanced model transparency by identifying daily reading duration, monthly book count, and selected digital-learning indicators as the strongest predictive contributors; however, these findings should be interpreted as predictive associations rather than causal determinants, and the observed behavioral patterns should not be treated as direct educational recommendations. Accordingly, the proposed framework may support preliminary screening of students who may require further attention in digitally mediated learning environments, rather than automated educational intervention. Before practical deployment, model predictions should be reviewed by educators and accompanied by fairness assessments, particularly when demographic variables are included, to minimize the risk of unintended bias. Nevertheless, several limitations should be acknowledged, including the use of self-reported questionnaire data, the possible conceptual overlap between reading-interest labels and reading-habit predictors, and the limited institutional context of the dataset. Future research should incorporate objective learning-management-system (LMS) activity logs, longitudinal behavioral observations, multicampus validation, fairness evaluation, and

intervention-effectiveness assessment to further improve model generalizability and practical applicability.

REFERENCES

[1] B. Albreiki, N. Zaki, and H. Alashwal, "A systematic literature review of student' performance prediction using machine learning techniques," *Educ. Sci. (Basel).*, vol. 11, no. 9, Sep. 2021, doi: 10.3390/educsci11090552.

[2] A. Namoun and A. Alshanqiti, "Predicting student performance using data mining and learning analytics techniques: A systematic literature review," Jan. 01, 2021, *MDPI AG*. doi: 10.3390/app11010237.

[3] U. Schiefele, E. Schaffner, J. Möller, and A. Wigfield, "Dimensions of Reading Motivation and Their Relation to Reading Behavior and Competence," *Read. Res. Q.*, vol. 47, no. 4, pp. 427–463, Oct. 2012, doi: 10.1002/RRQ.030.

[4] L. Chen, P. Chen, and Z. Lin, "Artificial Intelligence in Education: A Review," *IEEE Access*, vol. 8, pp. 75264–75278, 2020, doi: 10.1109/ACCESS.2020.2988510.

[5] N. Hidayat, L. Afuan, and H. R. Jannah, "Prescriptive Learning Analytics for Student Dropout: Integrating Temporal Velocity and Counterfactual Explanations in Longitudinal Data," *Journal of Computing Theories and Applications*, vol. 3, no. 4, pp. 627–646, May 2026, doi: 10.62411/jcta.15920.

[6] A. N. Firdaus and Y. R. Sipayung, "Stacking Ensemble Learning for University Student Dropout Prediction," *Journal of Information Systems and Informatics*, vol. 8, no. 1, pp. 456–477, Feb. 2026, doi: 10.63158/journalisi.v8i1.1403.

[7] H. Gu and Y. Zhang, "Efficient and Interpretable Machine Learning for Student Academic Outcome Prediction," *Mathematics*, vol. 14, no. 4, Feb. 2026, doi: 10.3390/math14040626.

[8] J. T. Guthrie and A. Wigfield, "Engagement and Motivation in Reading," in *Handbook of Reading Research*, vol. III, M. L. Kamil, P. B. Mosenthal, P. D. Pearson, and R. Barr, Eds., Mahwah, NJ: Lawrence Erlbaum Associates, 2000, pp. 403–422.

[9] P. Delgado, C. Vargas, R. Ackerman, and L. Salmerón, "Don't throw away your printed books: A meta-analysis on the effects of reading media on reading comprehension," *Educ. Res. Rev.*, vol. 25, pp. 23–38, Nov. 2018, doi: 10.1016/j.edurev.2018.09.003.

[10] H. Khosravi *et al.*, "Explainable Artificial Intelligence in education," *Computers and Education: Artificial Intelligence*, vol. 3, Jan. 2022, doi: 10.1016/j.caeai.2022.100074.

[11] W. Hidayatulloh, F. Mahardika, and D. I. Junaedi, "Explainable Artificial Intelligence-Based Model for Student Academic Performance Prediction," *Journal of Information System Exploration and Research*, vol. 4, no. 1, pp. 31–40, Feb. 2026, doi: 10.52465/joiser.v4i1.624.

[12] S. Wang and B. Luo, "Academic achievement prediction in higher education through interpretable modeling," *PLoS One*, vol. 19, no. 9, Sep. 2024, doi: 10.1371/journal.pone.0309838.

[13] OECD, *PISA 2018 Results (Volume II): Where All Students Can Succeed*, vol. II. Paris: OECD Publishing, 2019. doi: 10.1787/b5fd1b8f-en.

[14] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017, pp. 4765–4774.

[15] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA: Association for Computing Machinery (ACM), Jun. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[16] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.

[17] M. S. Khine, N. Ali, and O. Abu Khurma, "From Meals to Marks: Modeling the Impact of Family Involvement on Reading Performance with Counterfactual Explainable AI," *Educ. Sci. (Basel)*, vol. 15, no. 7, Jul. 2025, doi: 10.3390/educsci15070928.