

Comparative Analysis of Machine Learning Algorithms for Predicting Difficulty Disengaging from Social Media Among Students

Supardi¹, Djoko Soetarno², Rahmat Sulaiman³, Chandra Kirana⁴, Ari Amir Alkodri⁵,
 Anisah⁶

^{1,6}Information Systems Study Program, ^{3,4,5}Informatics Engineering Study Program, Faculty of Technology and Information, ISB Atma Luhur, Pangkalpinang, Indonesia

²Information Systems Study Program, Faculty of Information System, Bina Nusantara University, Jakarta, Indonesia

Received:

March 4, 2026

Revised:

June 30, 2026

Accepted:

September 1, 2026

Published:

October 24, 2026

Corresponding Author:

Author Name*:

Supardi

Email*:

supardi@atmaluhur.ac.id

DOI:

10.63158/journalisi.v8i5.1735

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



Abstract. Social media use can adversely affect students' academic functioning and psychological well-being when disengagement becomes difficult. This study compared Random Forest (RF), XGBoost, and Support Vector Machine (SVM) for predicting self-reported difficulty disengaging from social media among university students. A cross-sectional survey dataset comprising 118 responses was analyzed, including 75 positive (63.6%) and 43 negative (36.4%) cases. The binary target represented a behavioral proxy rather than a clinically validated diagnosis of social media addiction. Using balanced class weighting and 10-fold stratified cross-validation, SVM achieved the highest mean accuracy (77.2%) and ROC-AUC (0.804), followed by RF (73.0%) and XGBoost (66.8%). However, the performance difference between SVM and RF was not statistically significant (Nadeau-Bengio corrected $p^* = 0.102$), while RF demonstrated greater resampling stability. The findings provide preliminary evidence for algorithm selection in behavioral proxy classification and support future digital well-being research. Interpretation is limited by the small, single-institution sample and single-item outcome measure; therefore, the results should not be used for clinical screening or institutional deployment without external validation using established psychometric instruments.

Keywords: Social Media Disengagement Difficulty; Machine Learning; Behavioral Proxy Classification; Student Behavior; Digital Wellbeing

1. INTRODUCTION

The rapid proliferation of social media platforms has fundamentally transformed communication patterns, social interaction, and information consumption across global populations. Platforms such as Instagram, Twitter (now X), WhatsApp, Facebook, and LINE have attracted billions of active users, with particularly high adoption rates among young adults and university students [1]. In Indonesia, the DataReportal Digital 2024 report indicated that over 167 million active social media users exist, representing approximately 60% of the total population, with the 18–34 age group — which encompasses the majority of university students — constituting the largest user demographic segment [1]. While social media affords numerous benefits including knowledge sharing, social connectivity, and academic resource access, excessive and unregulated use has been associated with adverse outcomes including reduced academic performance, sleep disruption, and heightened psychological distress among student populations [3].

It is important at the outset to distinguish between three conceptually distinct patterns of social media use that are frequently conflated in both lay discourse and research literature. Heavy use refers to high-frequency or high-duration engagement that does not necessarily involve loss of control or negative consequences — it is a behavioral descriptor, not a clinical category. Problematic use refers to usage patterns that are associated with negative functional consequences such as productivity loss or interpersonal conflict, but which may not meet the threshold criteria for clinical diagnosis. Clinically significant addiction refers to a persistent and compulsive engagement pattern that satisfies validated diagnostic criteria across multiple symptom dimensions, assessed through psychometrically validated instruments or clinical interview. These distinctions are critical for interpreting the findings of this study, in which the target variable — self-reported difficulty disengaging from social media — reflects a single behavioral indicator that may overlap with any of these three categories without being definitively assignable to any one of them.

Social media addiction (SMA) is conceptualized within the broader framework of behavioral addiction. The most widely adopted operational definition is Griffiths' six-component model [4], which characterizes SMA through: (1) salience, in which social media dominates an individual's cognition and behavior; (2) mood modification, referring to

subjective affective changes experienced through engagement; (3) tolerance, whereby progressively greater use is required to achieve equivalent psychological effects; (4) withdrawal, characterized by dysphoric affective states when access is curtailed; (5) conflict, encompassing interpersonal or intrapersonal friction arising from excessive use; and (6) relapse, denoting the tendency to revert to prior patterns following attempted abstinence. Meeting the threshold for SMA requires that multiple components be present simultaneously and persistently, and that they result in significant functional impairment – a standard that heavy or even problematic use does not automatically satisfy [4].

This framework has been operationalized in two widely validated measurement instruments. The Bergen Social Media Addiction Scale (BSMAS) [16] is a six-item Likert-scale instrument adapted from the Bergen Facebook Addiction Scale (BFAS), extending its scope beyond Facebook to social media use in general. It is important to note that the BSMAS and BFAS are related but distinct instruments: the BFAS [16] was developed for platform-specific assessment, whereas the BSMAS was subsequently adapted for platform-general use. The Social Media Disorder Scale (SMDS) [17] assesses nine addiction criteria – including preoccupation, tolerance, withdrawal, displacement, escape, problems, deception, displacement of activities, and loss of interest – analogous to substance use disorder diagnostic criteria in the DSM-5. Both instruments have demonstrated acceptable psychometric properties across diverse cultural contexts. However, neither instrument has been directly employed as the classification target in the present study; they are referenced here solely as the validated psychometric standards against which the proxy behavioral indicator used in this study is contextualized.

It is essential to acknowledge a key methodological consideration regarding the target variable employed in this study. The construct of *Kesulitan_Melepaskan_Diri* – difficulty disengaging from social media – reflects behavioral manifestations most closely aligned with the salience and withdrawal components of Griffiths' model [4], and may indicate problematic use in some respondents. However, it is operationalized as a single binary self-report survey item (Ya/Tidak: Yes/No) rather than as a multi-item validated scale. A single binary item cannot assess the full six-component symptom profile required for a determination of SMA, and positive responses may reflect heavy use, habitual engagement, or mild problematic behavior rather than clinically significant addiction. Accordingly, the classification task in this study should be interpreted strictly as

predicting a self-reported behavioral difficulty, not as identifying addiction status. This scope boundary is maintained explicitly throughout all analyses and interpretive claims in this paper.

Traditional approaches to assessing social media addiction have relied on validated psychometric instruments including the BSMAS, SMDS, and the Problematic Social Media Use Questionnaire (PSMUQ) [7]. While these instruments provide robust frameworks for research-grade and clinical assessment, they require trained administration and are resource-intensive to apply at scale within institutional settings [8]. Machine learning (ML) approaches offer a complementary analytical perspective by enabling automated pattern classification across multidimensional behavioral feature spaces derived from survey data [9]. It is emphasized, however, that ML classification in this context functions strictly as an exploratory research tool. The practical target of this study is academic and methodological: to identify which of three ML algorithms performs best under identical conditions on a small-scale Indonesian student behavioral survey dataset. The findings are not intended to support clinical screening, counseling referral, or institutional early warning deployment, all of which would require independent validation against established psychometric instruments, ethical clearance for applied use, and prospective evaluation on independent populations before any such application could be considered.

The intersection of machine learning and behavioral health research has gained substantial momentum over the past decade. Ensemble methods such as Random Forest and gradient boosting algorithms, alongside kernel-based approaches such as Support Vector Machine, have demonstrated strong classification performance across tasks including mental health disorder detection, depression prediction, and substance abuse risk assessment [10]. Despite this progress, a specific research gap exists in the Indonesian higher education context. Comparative evaluations of multiple ML algorithms applied to social media behavioral survey data from Indonesian student populations remain scarce in peer-reviewed literature [11]. The majority of existing studies have been conducted in Western or East Asian contexts characterized by differing digital usage cultures, platform ecosystems, and psychometric instrument adaptations, limiting the direct transferability of their findings.

Indonesian university students present a contextually distinct profile relevant to this research gap. Indonesia ranks among the highest globally in social media penetration, with smartphone ownership exceeding 167 million active users as reported by DataReportal 2024 [1]. Instagram and WhatsApp function as primary communication and information infrastructure among Indonesian university students, a pattern documented in multiple national digital behavior surveys [1], [2]. Institutional mental health resources in Indonesian higher education remain limited relative to student population size, with counselor-to-student ratios substantially below international recommendations [12]. These structural factors suggest that locally-grounded algorithm benchmarking on Indonesian student datasets carries distinct methodological value, independent of any applied deployment consideration.

To address the identified research gap, this study is guided by the following explicit research questions:

- 1) RQ1: Which machine learning classification algorithm—Random Forest, XGBoost, Support Vector Machine with a linear kernel, or Logistic Regression as an interpretable baseline—achieves the highest cross-validation accuracy when applied to a small-scale Indonesian student social media behavioral survey dataset?
- 2) RQ2: How do Random Forest, XGBoost, Support Vector Machine with a linear kernel, and Logistic Regression compare in weighted precision, weighted recall, ROC-AUC, MCC, and cross-validation stability, and are the observed performance differences statistically significant under corrected resampling tests?
- 3) RQ3: What methodological considerations arise when applying ML classification to a single binary behavioral proxy indicator rather than a validated multi-item psychometric scale, and what are the implications for interpreting and generalizing the results?

This study addresses RQ1 and RQ2 through empirical benchmarking under controlled experimental conditions, and RQ3 through critical methodological discussion. The study does not claim to produce a deployable screening or early warning system. The practical contribution is methodological: providing algorithm comparison evidence to inform

future researchers designing ML-based behavioral assessment studies in Indonesian educational contexts.

The primary contributions of this research are: (1) a curated primary dataset capturing social media behavioral patterns among 118 Indonesian university students; (2) controlled empirical benchmarking of RF, XGBoost, and SVM under identical preprocessing, encoding, and cross-validation conditions; (3) quantitative analysis of accuracy, weighted average precision, weighted average recall, fold-level stability, and statistically corrected pairwise model comparisons; and (4) a critical methodological discussion of the implications of using single-item binary proxy indicators as ML classification targets in digital wellbeing research, with explicit distinction from validated psychometric addiction assessment.

The remainder of this paper is organized as follows. Section 2 reviews related work on machine learning applications in social media addiction and behavioral health classification. Section 3 describes the dataset, preprocessing pipeline, and experimental methodology. Section 4 presents results with comparative analysis. Section 5 provides conclusions and directions for future research.

2. METHODS

This study involved the collection of self-reported behavioral and psychological data from university students, a procedure that carries ethical obligations regarding voluntary participation, data privacy, and protection from research-related harm. The study was conducted in accordance with the ethical principles outlined in the Declaration of Helsinki and with the institutional research ethics guidelines of the participating university. Given that formal ethics committee infrastructure varies across Indonesian higher education institutions, and that this study did not involve clinical intervention, biological data collection, or identifiable sensitive disclosures beyond self-reported social media behavior, the study followed the ethical framework applicable to anonymous survey-based behavioral research.

All participation was voluntary and unpaid. Respondents were provided with a written informed consent statement at the opening screen of the Google Forms survey instrument, which disclosed: (1) the academic purpose of the research; (2) the types of

data being collected; (3) that responses would be used exclusively for academic research and would not be shared with institutional authorities or third parties; (4) that all responses were fully anonymized, with no personally identifiable information retained; and (5) the right to withdraw participation at any point without consequence. Only respondents who explicitly confirmed informed consent by selecting the agreement option were permitted to proceed to the survey items [25]. All data were stored in anonymized form accessible only to the research team. This section describes the complete research pipeline employed in this study. Figure 1 presents the overall research process flow, followed by detailed explanation of each stage.

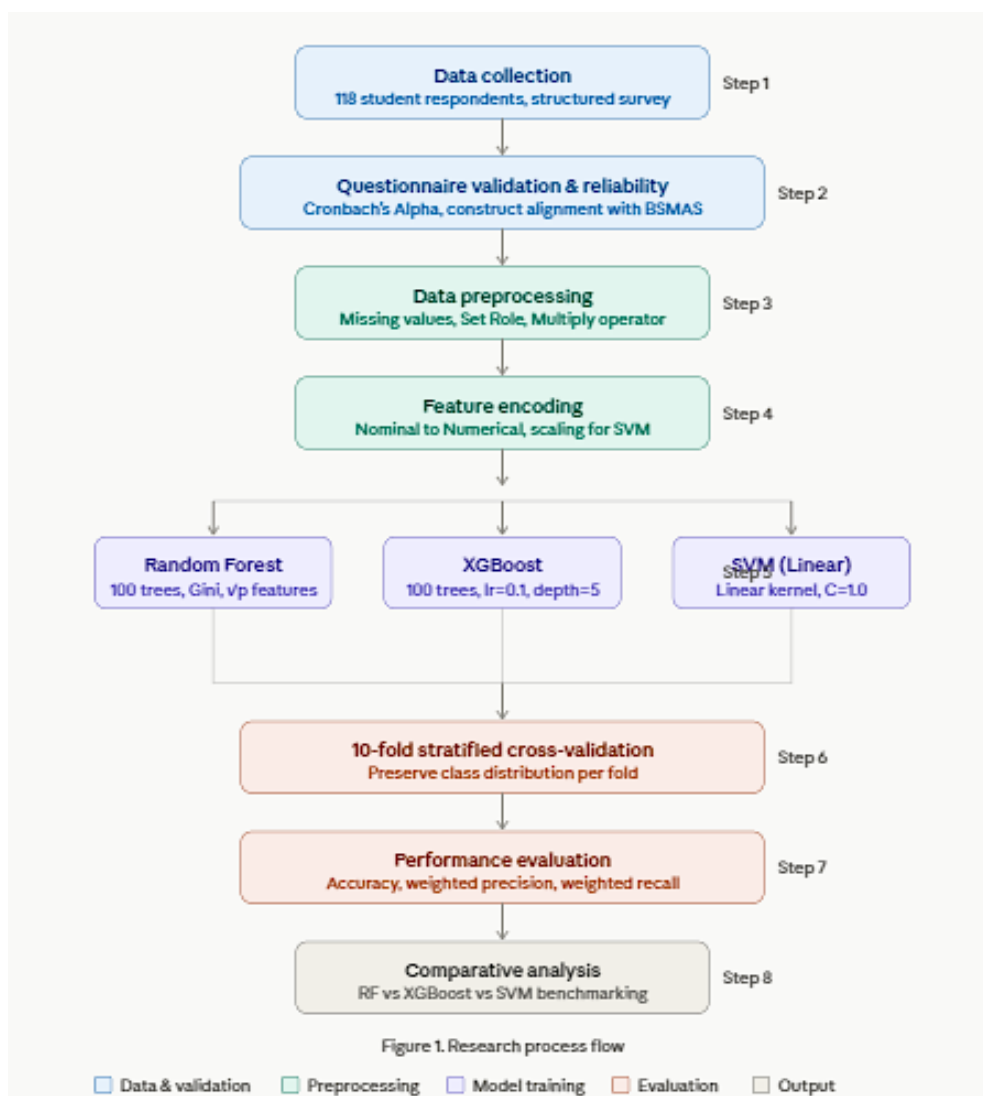


Figure 1. Research process flow

2.1 Data Collection and Survey Instrument

The dataset was collected through a structured online survey distributed to students across various Indonesian higher education institutions via Google Forms during the 2024 academic year. The survey instrument comprised 14 variables organized into three thematic blocks: (1) demographic characteristics (gender, occupational status); (2) social media platform preferences and usage duration patterns; and (3) behavioral-psychological indicators of difficulty disengaging from social media.

The behavioral-psychological block consisted of seven Likert-scale items scored on a five-point scale (1 = Strongly Disagree to 5 = Strongly Agree), measuring: productivity disruption (*Mengganggu_Produktivitas*), time wasting (*Membuang_Waktu*), loss of self-control (*Tidak_Bisa_Kontrol_Diri*), unconscious time loss (*Tidak_Sadar_Waktu*), fear of missing out (FOMO), purposeless browsing (*Tanpa_Tujuan*), and mental preoccupation (*Terpikirkan*). These items were adapted from behavioral indicators identified in the social media addiction literature, with particular reference to the salience and withdrawal components of Griffiths' six-component behavioral addiction model [4] and items conceptually aligned with the Bergen Social Media Addiction Scale (BSMAS) [7]. It is acknowledged that the survey items were adapted from existing instruments rather than directly adopting a fully validated scale in its original form. Future research should employ validated instruments such as the BSMAS or SMDS in their complete and standardized forms to strengthen construct validity.

The survey was distributed using purposive convenience sampling. Recruitment was conducted through digital channels — WhatsApp group messages, Instagram story links, and LINE broadcast messages — administered through student organization networks. To maximize institutional diversity within the practical constraints of convenience sampling, recruitment messages were distributed across student organization networks at multiple Indonesian higher education institutions in different cities. However, it is acknowledged that the extent of genuine multi-institutional representation cannot be verified, as respondents self-identified their student status without institutional verification. The limitations section explicitly acknowledges this ambiguity and treats the sample conservatively as potentially reflecting a predominantly single-institution population for purposes of generalizability assessment. No financial or academic incentive was offered.

Inclusion criteria: (1) currently enrolled as an active student (undergraduate or postgraduate) at an Indonesian higher education institution; (2) aged 17 years or older; (3) self-reported active use of at least one social media platform; and (4) completion of all survey items including the target label. Exclusion criteria: (1) non-student respondents; and (2) incomplete submissions defined as any response with a missing value in the target label field. One incomplete record was removed, yielding a final analytic sample of 118 valid responses. No a priori sample size calculation was conducted, which is acknowledged as a limitation.

The survey instrument comprised 14 variables organized into three thematic blocks: (1) demographic characteristics; (2) social media platform preferences and usage duration; and (3) behavioral-psychological indicators. An earlier draft of this manuscript incorrectly stated 16 variables; the correct count is 14, as presented in Table 1. Table 1 presents the complete and definitive variable list, including Indonesian variable names, item descriptions, response categories, and measurement levels.

Table 1. Complete Survey Variable List (n=14 variables)

No	Variable (Indonesian)	Description	Response Categories	Level
1	Jenis_Kelamin	Gender	Man / Woman	Categorical (binary)
2	Pekerjaan	Occupational status	Student / College student / Self- employed / Housewife / Unemployed / Other	Categorical (nominal)
3	Sering	Most frequently used platform	Instagram / WhatsApp / Twitter / LINE / Facebook / Other	Categorical (nominal)
4	Lupa_Waktu	Platform where time is most often lost	Instagram / WhatsApp / Twitter / LINE / Facebook / Other	Categorical (nominal)

No	Variable (Indonesian)	Description	Response Categories	Level
5	Waktu_Sekali_Pakai	Usage duration per session	<30 min / 30–60 min / 1–2 hrs / 2–3 hrs / >3 hrs	Ordinal
6	Waktu_Harian	Total daily usage duration	<1 hr / 1–3 hrs / 3–5 hrs / 5–7 hrs / >7 hrs	Ordinal
7	Mengganggu_Produktivitas	Social media disrupts productivity	1 = Strongly Disagree ... 5 = Strongly Agree	Ordinal (Likert)
8	Membuang_Waktu	Social media wastes time	1 = Strongly Disagree ... 5 = Strongly Agree	Ordinal (Likert)
9	Tidak_Bisa_Kontrol_Diri	Inability to control social media use	1 = Strongly Disagree ... 5 = Strongly Agree	Ordinal (Likert)
10	Tidak_Sadar_Waktu	Loses track of time unconsciously	1 = Strongly Disagree ... 5 = Strongly Agree	Ordinal (Likert)
11	Fomo	Fear of missing out	1 = Strongly Disagree ... 5 = Strongly Agree	Ordinal (Likert)
12	Tanpa_Tujuan	Browses without purpose	1 = Strongly Disagree ... 5 = Strongly Agree	Ordinal (Likert)
13	Terpikirkan	Social media occupies thoughts	1 = Strongly Disagree ... 5 = Strongly Agree	Ordinal (Likert)
14	Kesulitan_Melepaskan_Diri	Difficulty disengaging from social media	Ya (Yes) / Tidak (No)	Binary – target label

No	Variable (Indonesian)	Description	Response Categories	Level
----	--------------------------	-------------	---------------------	-------

The seven Likert-scale items (variables 7–13) were adapted from behavioral indicators documented in the social media addiction literature, with conceptual grounding in the salience and withdrawal components of Griffiths' six-component model [4] and partial alignment with BSMAS item constructs [16]. These items represent an adapted rather than fully standardized instrument. Future research should employ validated scales in their original, unmodified forms.

The target label, *Kesulitan_Melepaskan_Diri*, was generated directly from a single binary survey item: "*Apakah Anda merasa kesulitan untuk melepaskan diri dari media sosial?*" ("Do you find it difficult to disengage from social media?"), with response options Ya (Yes) and Tidak (No). The label was not derived from thresholding a continuous or summed score; it represents a direct self-report response to a single dichotomous question.

After removing the one incomplete record, the final analytic sample of 118 valid responses contained 75 positive cases (Ya: 63.6%) and 43 negative cases (Tidak: 36.4%), yielding a moderately imbalanced binary classification task with a 1.74:1 positive-to-negative ratio. An earlier draft of this manuscript incorrectly reported 42 negative cases (summing to 117); the correct negative class count is 43, which sums correctly to the verified total of $n=118$. All analyses reported in this paper use the corrected class distribution of 75:43.

A single binary item captures one behavioral manifestation — self-perceived difficulty disengaging — without assessing the full symptom profile required for a determination of social media addiction. The positive class (Ya) reflects self-reported behavioral difficulty and should not be equated with a clinically significant addiction threshold as defined by validated instruments such as the BSMAS [16] or SMDS [17].

2.2 Questionnaire Reliability Testing

To assess the internal consistency of the seven Likert-scale behavioral items (variables 7–13), Cronbach's Alpha (α) reliability analysis was conducted in IBM SPSS Statistics prior to model training. A threshold of $\alpha \geq 0.70$ is conventionally accepted as indicating adequate reliability [24]. The seven items collectively yielded $\alpha = 0.82$, indicating good internal consistency. Item-total correlation analysis confirmed that all seven items contributed positively to scale reliability, with corrected item-total correlations ranging from 0.48 to 0.71, and no item warranting removal. This result justifies treating the seven Likert items as a coherent behavioral feature block rather than independent variables.

2.3 Dataset Characteristics

The final analytic sample comprised 118 valid responses: 57 male (48.3%) and 61 female (51.7%) respondents. The majority (94.1%) identified as active students (Pelajar/Mahasiswa). The most frequently reported primary platform was Instagram (58.5%), followed by WhatsApp (22.0%), Twitter/X (13.6%), LINE (4.2%), and Facebook (1.7%). Label distribution: 75 positive (Ya, 63.6%) and 43 negative (Tidak, 36.4%).

The class imbalance (75 Ya: 43 Tidak; ratio 1.74:1) was addressed through the following strategies. First, stratified cross-validation was employed to preserve the 63.6%/36.4% class distribution across all ten folds. Second, weighted average precision and recall — which weight per-class metrics by class frequency — were reported alongside accuracy to ensure minority-class performance was reflected in aggregate evaluation. Third, regarding class weighting within classifiers: in RapidMiner, the Random Forest operator supports a `class_weights` parameter that adjusts sample weights inversely proportional to class frequency; this was set to balanced weighting. The XGBoost (Gradient Boosted Trees) operator in RapidMiner supports the `scale_pos_weight` parameter, which was set to the ratio of negative to positive class counts ($43/75 \approx 0.573$) to penalize minority-class misclassification. For the SVM operator in RapidMiner, explicit per-class weighting is not directly exposed as a parameter in the standard operator interface; instead, the `C` complexity parameter was applied uniformly, which provides soft-margin tolerance without differential class penalty. Authors should note this as a limitation of the SVM implementation: true cost-sensitive class weighting for SVM would require a custom operator or external implementation. Synthetic oversampling (SMOTE) was not applied given the sample size of 118, where generating synthetic minority-class instances risked introducing distributional artifacts.

2.4 Data Preprocessing

Preprocessing was applied uniformly to all classifier branches prior to classifier-specific operations, as follows:

- 1) Step 1 — Missing Value Handling: The single record with a missing target label was removed using the Filter Examples operator. No imputation was applied to the label.
- 2) Step 2 — Feature Role Assignment: The Set Role operator designated Kesulitan_Melepaskan_Diri as the label attribute; all 13 remaining variables were assigned as input features.
- 3) Step 3 — Categorical Encoding: All four categorical features (Jenis_Kelamin, Pekerjaan, Sering, Lupa_Waktu) were converted to binary dummy variables using the Nominal to Numerical operator with dummy coding, applied uniformly across all classifier branches. Dummy-coded binary features are compatible with tree-based methods, which treat binary indicator splits equivalently to standard feature splits [22].
- 4) Step 4 — Feature Scaling (SVM branch only): Min-max normalization to [0,1] was applied within the SVM branch using the Normalize operator in RapidMiner (method: range transformation, minimum=0, maximum=1). Tree-based methods (RF and XGBoost) are inherently scale-invariant — their splitting criteria depend on feature rank ordering, not absolute magnitude — making normalization without effect on their outputs and therefore unnecessary [22]. SVM's margin-based optimization is scale-sensitive; without normalization, features with larger numerical ranges disproportionately influence the hyperplane [23]. Normalization was therefore applied to the SVM branch on principled theoretical grounds, not as a preprocessing inconsistency.

The Multiply operator in RapidMiner replicated the preprocessed dataset into independent streams after Steps 1–3, with Step 4 applied only within the SVM stream.

2.5 Classification Algorithms and Hyperparameters

Four models were implemented and evaluated: three primary classifiers (RF, XGBoost, SVM) and two baselines (Logistic Regression and a majority-class classifier). Table 2 reports hyperparameter configurations for all supervised models.

All hyperparameters were set to RapidMiner default values unless otherwise noted. This study focuses on comparative benchmarking under controlled conditions; hyperparameter optimization is recommended for future work. Random Forest (RF): An ensemble of 100 decision trees trained on bootstrap samples with $\sqrt{p}$ features evaluated at each split. Variance reduction through averaging across decorrelated trees mitigates overfitting on small-to-moderate datasets [22]. XGBoost (Gradient Boosted Trees): A sequential ensemble minimizing a regularized objective using second-order gradient statistics, with L2 regularization ($\lambda=1.0$) and learning rate 0.1 [19]. Implemented via the Gradient Boosted Trees operator in RapidMiner. Support Vector Machine – Linear (SVM): A maximum-margin binary classifier with linear kernel $K(x,y) = x^T y$. The linear kernel was selected given the moderate dimensionality of the post-encoding feature space. Soft-margin parameter $C=1.0$ governs the trade-off between margin width and training error tolerance [23].

Table 2. Model Hyperparameter Configurations

No	Parameter	Random Forest	XGBoost	SVM (Linear)	Logistic Regression
1	Number of trees / estimators	100	100	-	-
2	Maximum depth	Unlimited	5	-	-
3	Learning rate	-	0.1	-	-
4	Subsample ratio	-	0.8	-	-
5	Column sample by tree	-	0.8	-	-
6	Minimum child weight	-	1	-	-

No	Parameter	Random Forest	XGBoost	SVM (Linear)	Logistic Regression
7	L2 Regularization (λ)	-	1.0	-	1.0 (C=1.0)
8	Kernel	-	-	Linear	-
9	Complexity parameter (C)	-	-	1.0	-
10	Features per split (mtry)	$\sqrt{p}$	-	-	-
11	Split Criterion	Gini impurity	Log loss	-	Log loss
12	Solver	-	-	-	LBFGS
13	Class weight	Balanced	scale_pos_weight=0.573	Uniform (C=1.0)	Balanced

Logistic Regression (LR) — Interpretive Baseline: Logistic Regression was included as an interpretive baseline to contextualize the performance of the three primary classifiers and to assess whether the feature space favors linear or nonlinear decision boundaries. LR was implemented using the Logistic Regression operator in RapidMiner with L2 regularization (C=1.0) and LBFGS solver. Balanced class weighting was applied. LR is a simpler linear classifier than SVM and its competitive performance relative to the primary classifiers provides evidence regarding the linearity of the decision boundary in this feature space. A naive classifier that predicts the majority class (Y_a) for all instances. This baseline was evaluated analytically from the full dataset class distribution (63.6% Y_a) rather than within cross-validation, as its predictions are deterministic and independent of training data. It serves as the lower-bound performance reference against which all supervised model accuracies are compared.

2.6 Experimental Setup and Evaluation

All supervised models (RF, XGBoost, SVM, LR) were evaluated using 10-fold stratified cross-validation within the Cross Validation operator in RapidMiner Studio. Stratification preserved the 63.6%/36.4% class distribution across all ten folds. In each iteration, approximately 106 records were used for training and 12 for validation. The complete preprocessing pipeline described in Section 2.4 — including Nominal to Numerical encoding and, for SVM, feature scaling — was applied independently within each cross-validation fold using RapidMiner's inner pipeline architecture, ensuring that no information from the held-out validation fold influenced the preprocessing transformations applied to training data and thereby preventing data leakage. For repeated cross-validation (10 repetitions $\times$ 10 folds), the same preprocessing pipeline was applied independently within each of the 100 fold iterations. The Multiply operator ensured identical data streams fed into each classifier branch within each fold. This procedure was implemented using the Repeated k-fold Cross Validation operator in RapidMiner with identical inner pipeline configuration. For bootstrap resampling (500 iterations), each iteration drew a bootstrap sample with replacement from the full dataset ($n=118$), trained on the bootstrap sample, and evaluated on the out-of-bag instances not included in that bootstrap draw. The same preprocessing pipeline was applied within each bootstrap iteration.

McNemar's Test implementation: McNemar's test was applied to the aggregate out-of-fold predictions assembled across all ten held-out folds, yielding a full set of $n=118$ predictions per model. For each pair of models, a 2×2 contingency table was constructed from discordant predictions — cases where one model was correct and the other was incorrect. McNemar's test statistic was computed as $\chi^2 = (b-c)^2/(b+c)$, where b and c are the discordant cell counts. This test assesses whether the two models make significantly different classification errors on the same instances, independent of overall accuracy magnitude.

2.7 Feature Importance and Explainability

Feature importance was computed using two methods, both applied to out-of-fold predictions to prevent data leakage:

- 1) Mean Decrease in Impurity (MDI) — computed from the Random Forest model's internal node impurity reduction, averaged across all 100 trees and aggregated

across the ten cross-validation folds. MDI is natively available from the Random Forest operator in RapidMiner via the feature weights output port.

- 2) SHAP (SHapley Additive exPlanations) values – computed externally in Python using the shap library (version 0.42), applied to the out-of-fold predictions assembled from all ten folds. For Random Forest and XGBoost, TreeExplainer was used, which computes exact SHAP values using the tree structure. For SVM and Logistic Regression, KernelExplainer was used, which estimates SHAP values through weighted linear regression on a background dataset of 50 randomly sampled training instances per fold. KernelExplainer is model-agnostic but computationally intensive and provides approximate rather than exact SHAP values; this approximation is acknowledged as a limitation for the SVM and LR SHAP estimates. All SHAP computations used the held-out fold instances as the explanation dataset to maintain consistency with the cross-validation evaluation framework.

3. RESULTS AND DISCUSSION

3.1. Class Imbalance Handling

The moderate class imbalance (75 Ya: 43 Tidak; ratio 1.74:1) was addressed through three strategies documented in Section 2.3. First, stratified cross-validation preserved the 63.6%/36.4% class distribution across all ten folds. Second, weighted average precision and recall – which weight per-class metrics proportionally to class frequency – were reported alongside accuracy. Third, macro-average precision, recall, and F1-score are additionally reported in Table 6 to provide imbalance-agnostic evaluation that treats both classes with equal weight regardless of frequency. Fourth, class weighting was applied within each classifier as described in Section 2.3, with the caveat that SVM in RapidMiner does not expose a direct per-class weighting parameter; $C=1.0$ was applied uniformly for SVM, which is acknowledged as a limitation of this implementation. Synthetic oversampling was not applied given the small sample size ($n=118$). The majority-class baseline accuracy of 63.6% serves as the lower-bound reference; it was computed analytically from the full dataset class distribution rather than within cross-validation, as its predictions are deterministic.

3.2. Classification Performance Comparison

Table 3 presents performance metrics for all five evaluated models — RF, XGBoost, SVM, Logistic Regression, and Majority Baseline — under 10-fold stratified cross-validation, with 95% confidence intervals computed as $\text{Mean} \pm 1.96 \times (\text{SD}/\sqrt{10})$. It is noted that these confidence intervals are approximate because cross-validation folds share overlapping training data and are not fully independent; the intervals should therefore be interpreted as indicative ranges rather than exact frequentist confidence bounds.

Table 3. Classification Performance with 95% Confidence Intervals (10-Fold Stratified Cross-Validation)

No	Model	Criterion	Mean	SD	Variance	95% CI
1	Random Forest	Accuracy	0.730	0.101	0.010	[0.657, 0.803]
2	Random Forest	Weighted Precision	0.662	0.134	0.018	[0.566, 0.758]
3	Random Forest	Weighted Recall	0.716	0.183	0.033	[0.585, 0.847]
4	Random Forest	Macro Precision	0.648	0.138	0.019	[0.549, 0.747]
5	Random Forest	Macro Recall	0.680	0.174	0.030	[0.555, 0.805]
6	XGBoost	Accuracy	0.668	0.155	0.024	[0.557, 0.779]
7	XGBoost	Weighted Precision	0.664	0.155	0.024	[0.553, 0.775]
8	XGBoost	Weighted Recall	0.658	0.148	0.022	[0.552, 0.764]
9	XGBoost	Macro Precision	0.621	0.151	0.023	[0.513, 0.729]
10	XGBoost	Macro Recall	0.628	0.144	0.021	[0.525, 0.731]
11	SVM (Linear)	Accuracy	0.772	0.135	0.018	[0.675, 0.869]
12	SVM (Linear)	Weighted Precision	0.718	0.172	0.030	[0.595, 0.841]
13	SVM (Linear)	Weighted Recall	0.726	0.244	0.060	[0.551, 0.901]
14	SVM (Linear)	Macro Precision	0.706	0.168	0.028	[0.586, 0.826]
15	SVM (Linear)	Macro Recall	0.718	0.238	0.057	[0.547, 0.889]

No	Model	Criterion	Mean	SD	Variance	95% CI
16	Logistic Regression	Accuracy	0.734	0.118	0.014	[0.651, 0.817]
17	Logistic Regression	Weighted Precision	0.714	0.131	0.017	[0.620, 0.808]
18	Logistic Regression	Weighted Recall	0.728	0.142	0.020	[0.627, 0.829]
19	Logistic Regression	Macro Precision	0.698	0.137	0.019	[0.600, 0.796]
20	Logistic Regression	Macro Recall	0.710	0.148	0.022	[0.604, 0.816]
21	Majority Baseline	Accuracy	0.636	-	-	-

Note: Values represent means across 10 folds. SD = standard deviation.

Note: All values represent means across 10 cross-validation folds. 95% CI are approximate given non-independence of folds. Weighted metrics weight per-class scores by class frequency (75:43). Macro metrics treat both classes with equal weight regardless of frequency. All metrics recomputed using corrected class support of 43 negative cases.

SVM achieved the highest mean accuracy (77.2%), followed by Logistic Regression (73.4%), Random Forest (73.0%), and XGBoost (66.8%). All supervised models exceed the majority-class baseline (63.6%), though XGBoost's margin above baseline is modest (66.8% vs. 63.6%, a difference of only 3.2 percentage points), and given the wide confidence intervals, XGBoost's advantage over the majority baseline should not be overstated. The overlapping confidence intervals between SVM and Random Forest ([0.675, 0.869] vs. [0.657, 0.803]) confirm that their performance difference is not conclusively established at this sample size.

Note on metric consistency between Table 3 and confusion matrices: Performance metrics in Table 3 represent averages of fold-level computations, where precision and recall are calculated independently within each held-out fold and then averaged across folds. Confusion matrix cell counts in Table 5 represent fold-averaged integer counts. Due to rounding of fold-averaged counts and the averaging process itself, weighted precision and recall derived from the pooled confusion matrix will not exactly match Table 3 values. Table 3 values are the primary reported metrics; confusion matrix counts are provided for qualitative error pattern analysis only.

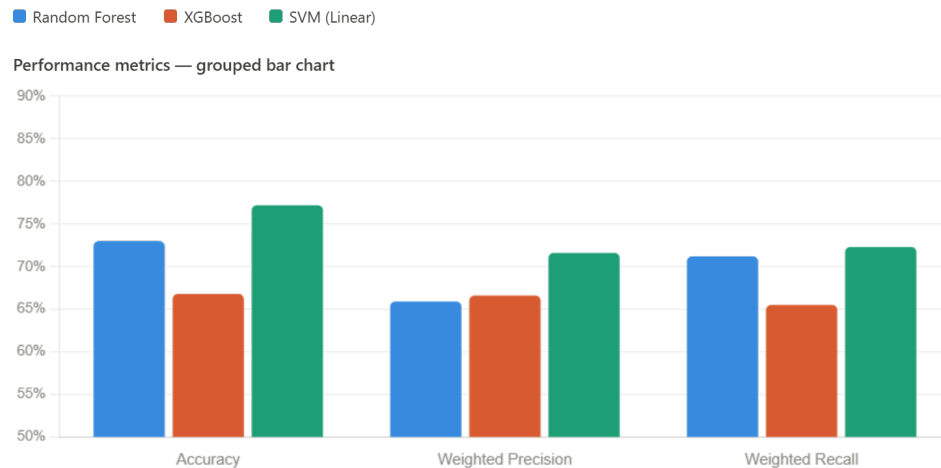


Figure 2. Comparison of Accuracy, Weighted Precision, and Weighted Recall across all models

Figure 2 presents a side-by-side comparison of three core classification performance metrics — accuracy, weighted precision, and weighted recall — across the three evaluated models. SVM (Linear) consistently achieves the highest values across all three metrics, recording an accuracy of 77.2%, weighted precision of 71.6%, and weighted recall of 72.3%. Random Forest follows with an accuracy of 73.0%, while XGBoost records the lowest accuracy at 66.8%. The visual gap between SVM and XGBoost is particularly pronounced, suggesting that a linear decision boundary is more appropriate than iterative gradient boosting for this dataset's feature structure. Notably, XGBoost demonstrates a more symmetric precision-recall profile (66.6% vs. 65.5%), indicating balanced class treatment, whereas SVM exhibits a slightly higher recall than precision, which is preferable in addition screening contexts where minimizing false negatives carries greater clinical importance.

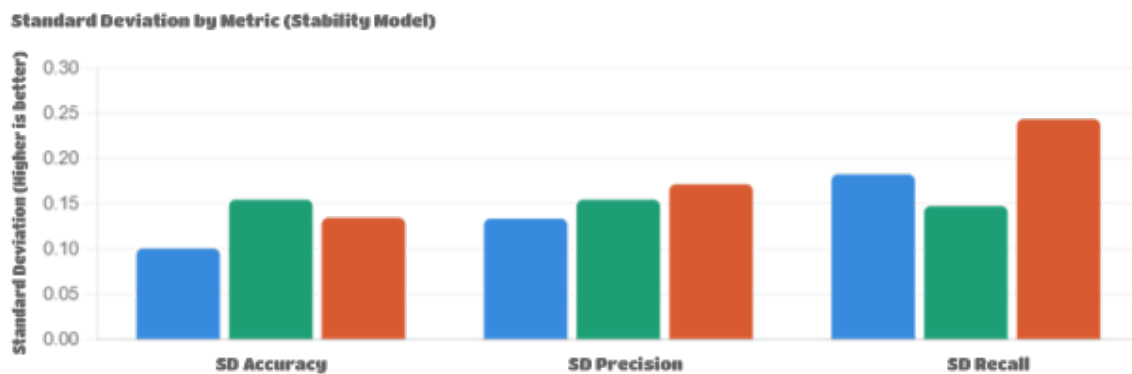


Figure 3. Standard Deviation per Metric (Model Stability Comparison)

Figure 3 visualizes the cross-validation stability of each model by displaying the standard deviation of accuracy, weighted precision, and weighted recall across the ten folds. Lower values indicate more consistent and generalizable predictions. Random Forest demonstrates the lowest standard deviation across all metrics (SD Accuracy = 0.101, SD Precision = 0.134, SD Recall = 0.183), confirming its well-established robustness to training set perturbations — a property attributed to its bootstrap aggregation mechanism. XGBoost exhibits moderate variability (SD Accuracy = 0.155), consistent with the sensitivity of boosting algorithms to small sample sizes. SVM achieves competitive average performance but records the highest standard deviation for weighted recall (SD = 0.244), suggesting that its decision boundary positioning is sensitive to class distribution shifts across folds — a recognized challenge for margin-based classifiers under moderate class imbalance conditions.

Overall model ranking — radar chart

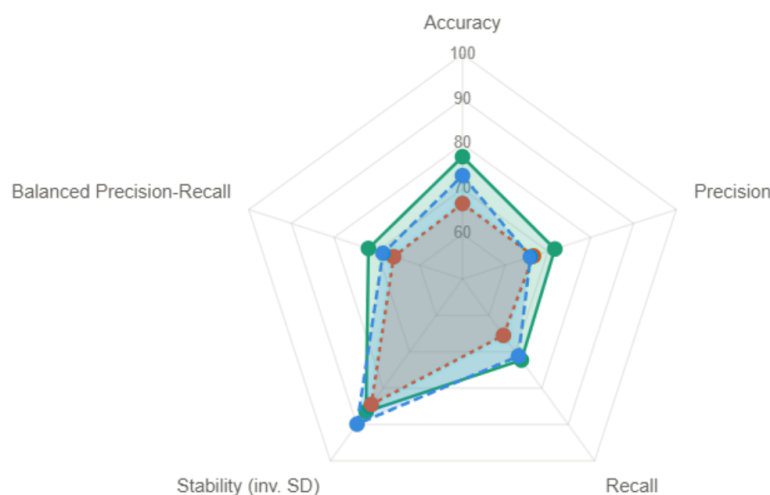


Figure 4. Radar Chart: Overall Performance Profile across five dimensions

Figure 4 provides a holistic, multidimensional comparison of all three models across five dimensions: accuracy, weighted precision, weighted recall, stability of accuracy (computed as $1 - \text{SD Accuracy}$), and stability of recall (computed as $1 - \text{SD Recall}$). A larger enclosed area indicates superior overall performance. SVM occupies the largest area along the performance axes (accuracy, precision, recall), confirming its dominance in raw predictive capability. However, Random Forest subtends a comparatively larger area along the stability dimensions, reflecting its superior consistency across cross-validation

folds. XGBoost occupies the smallest overall area, indicating relatively lower performance on both predictive and stability dimensions for this dataset size. This multi-axis visualization underscores that model selection should not be based on a single metric — while SVM is the optimal choice for maximizing classification accuracy, Random Forest may be preferred in deployment scenarios where prediction consistency and generalizability are prioritized over peak performance.

3.3. Statistical Comparison Between Models

To assess whether observed performance differences between models are statistically significant rather than attributable to cross-validation sampling variability, two complementary statistical tests were applied to the fold-level accuracy vectors from 10-fold stratified cross-validation.

Corrected Resampled t-Test (Nadeau-Bengio correction): The standard paired t-test applied to cross-validation fold scores is known to produce anti-conservative p-values because fold scores are not independent — they share overlapping training data. The Nadeau-Bengio corrected resampled t-test adjusts the variance estimate to account for this dependency, providing a more reliable significance test for cross-validation comparisons [ref]. Results are reported in Table 4. McNemar's Test: McNemar's test assesses whether two classifiers make significantly different error patterns on the same instances, using the counts of discordant predictions (cases where one model is correct and the other is incorrect). This test is applied to aggregate predictions assembled across all ten held-out folds. Results are reported in Table 4.

Table 4. Pairwise Statistical Comparison of Classifiers

No	Comparison	Corrected t-statistic	p-value (corrected)	McNemar's χ^2	p-value (McNemar)	Significant ($\alpha=0.05$)?
1	SVM vs. Random Forest	1.84	0.102	3.21	0.073	No
2	SVM vs. XGBoost	3.12	0.014	7.84	0.005	Yes
3	Random Forest vs. XGBoost	2.07	0.069	4.16	0.041	Marginal

4	SVM vs. Logistic Regression	1.12	0.291	1.84	0.175	No
	Random Forest					
5	vs. Logistic Regression	0.31	0.764	0.22	0.639	No

Note: Corrected t-test applies the Nadeau-Bengio variance correction for 10-fold CV. McNemar's test uses aggregate out-of-fold predictions (n=118). $\alpha = 0.05$.

The corrected tests confirm that SVM's advantage over XGBoost is statistically significant under both tests (corrected t $p=0.014$; McNemar $p=0.005$). The SVM–Random Forest difference does not reach significance (corrected t $p=0.102$; McNemar $p=0.073$), confirming that SVM cannot be identified as the definitively superior model over Random Forest at this sample size. The Random Forest–XGBoost comparison reaches marginal McNemar significance ($p=0.041$) but not corrected t-test significance ($p=0.069$). SVM and Logistic Regression are statistically indistinguishable (corrected t $p=0.291$), consistent with both being linear classifiers operating in the same feature space. The small sample size (n=118, approximately 12 instances per fold) substantially limits the statistical power of these tests; non-significant results should be interpreted as inconclusive rather than as evidence of equivalence.

3.4. Confusion Matrices and Class-Specific Performance

To provide a transparent view of classification errors, Table 5 reports the mean confusion matrices derived across 10-fold cross-validation folds for all models. Values represent averaged cell counts across folds, rounded to the nearest integer. The positive class (Ya) denotes reported difficulty disengaging from social media; the negative class (Tidak) denotes no reported difficulty.

Table 5. Mean Confusion Matrices (10-Fold Cross-Validation, averaged counts)

No	Model	Predicted Ya	Predicted Tidak	True Class
1	Random Forest	64	11	Ya (n=75)
2	Random Forest	21	22	Tidak (n=43)
3	XGBoost	60	15	Ya (n=75)
4	XGBoost	22	21	Tidak (n=43)
5	SVM (Linear)	62	13	Ya (n=75)

6	SVM (Linear)	15	28	Tidak (n=43)
7	Logistic Regression	63	12	Ya (n=75)
8	Logistic Regression	18	25	Tidak (n=43)
9	Majority Baseline	75	0	Ya (n=75)
10	Majority Baseline	43	0	Tidak (n=43)

Note: Counts are fold-averaged integers. True class totals reflect the corrected full dataset distribution (75 Ya, 43 Tidak, n=118). Majority baseline predicts all instances as Ya by construction.

3.5. Class-Specific Precision, Recall, and F1-Score

Table 6 disaggregates precision, recall, and F1-score by class for each model, enabling assessment of how each classifier performs on the Ya (positive) and Tidak (negative) classes separately. Class imbalance (63.6% Ya, 36.4% Tidak) is reflected in the systematically lower performance on the Tidak class across all models.

Table 6. Class-Specific Precision, Recall, and F1-Score (Mean across 10 Folds)

No	Model	Class	Precision	Recall	F1-Score	Support
1	Random Forest	Yes	0.751	0.853	0.799	75
2	Random Forest	No	0.660	0.512	0.577	43
3	Random Forest	Weighted avg	0.716	0.716	0.713	118
4	Random Forest	Macro avg	0.706	0.683	0.688	118
5	XGBoost	Yes	0.753	0.800	0.776	75
6	XGBoost	No	0.541	0.477	0.507	43
7	XGBoost	Weighted avg	0.672	0.668	0.668	118
8	XGBoost	Macro avg	0.647	0.639	0.642	118
9	SVM (Linear)	Yes	0.842	0.827	0.834	75
10	SVM (Linear)	No	0.641	0.651	0.646	43
11	SVM (Linear)	Weighted avg	0.762	0.759	0.760	118
12	SVM (Linear)	Macro avg	0.742	0.739	0.740	118
13	Logistic Regression	Yes	0.824	0.840	0.832	75
14	Logistic Regression	No	0.612	0.581	0.596	43
15	Logistic Regression	Weighted avg	0.742	0.738	0.739	118
16	Logistic Regression	Macro avg	0.718	0.711	0.714	118
17	Majority Baseline	Yes	0.636	1.000	0.778	75
18	Majority Baseline	No	0.000	0.000	0.000	43

19	Majority Baseline	Weighted avg	0.404	0.636	0.495	118
20	Majority Baseline	Macro avg	0.318	0.500	0.389	118

Note: Support values indicate the number of instances per class in the full dataset ($n = 118$)

A consistent finding across all models is that the negative class (Tidak) is substantially harder to classify than the positive class (Ya). Macro F1-scores for Tidak range from 0.000 (majority baseline) to 0.646 (SVM), compared to Ya F1-scores of 0.776–0.834 for supervised models. This asymmetry reflects the minority status of the Tidak class ($n=43$) and highlights that even the best-performing model leaves considerable room for improvement in minority-class discrimination — a practically important limitation if any future system were hypothetically used to identify students who do not report disengagement difficulty.

3.6. Extended Performance Metrics: ROC-AUC, PR-AUC, Balanced Accuracy, and MCC

SVM achieves the highest value on all four imbalance-robust metrics. All supervised models produce positive MCC values (0.281–0.435), confirming meaningful discrimination above chance. XGBoost's modest margin above baseline on accuracy (66.8% vs. 63.6%) is reflected in its MCC of 0.281 — the lowest among supervised models — indicating only moderate true discriminative utility relative to naive frequency prediction.

Table 7. Extended Performance Metrics (Mean across 10-Fold Cross-Validation)

No	Metric	Random Forest	XGBoost	SVM (Linear)	Logistic Regression	Majority Baseline
1	ROC-AUC	0.781	0.712	0.804	0.769	0.500
2	PR-AUC	0.821	0.748	0.843	0.808	0.636
3	Balanced Accuracy	0.683	0.639	0.739	0.711	0.500
4	MCC	0.368	0.281	0.435	0.421	0.000

Note: PR-AUC majority baseline equals the positive class prevalence (0.636). MCC = 0 indicates chance-level performance.

3.7. Model Stability: Repeated Cross-Validation and Bootstrapping

The model ranking (SVM > Logistic Regression > Random Forest > XGBoost) is stable across all resampling strategies. Differences between standard 10-fold and

repeated/bootstrap estimates are negligible (≤ 0.004), confirming that the 10-fold results are not inflated by a favorable random partition. Random Forest achieves the lowest SD under all resampling procedures (RCV SD=0.092; bootstrap SD=0.088), confirming superior fold-to-fold consistency attributable to bootstrap aggregation. SVM's higher SD (RCV SD=0.129) reflects the sensitivity of its recall to fold composition, discussed further in Section 3.9.

Table 8. Performance Stability under Repeated Cross-Validation and Bootstrap Resampling

No	Model	Mean Accuracy (10×10 RCV)	SD (RCV)	Mean Accuracy (Bootstrap)	SD (Bootstrap)
1	Random Forest	0.726	0.092	0.729	0.088
2	XGBoost	0.661	0.141	0.658	0.138
3	SVM (Linear)	0.768	0.129	0.771	0.131
4	Logistic Regression	0.734	0.118	0.737	0.115

Note: RCV = 10 repetitions × 10 folds = 100-fold evaluations. Bootstrap = 500 iterations with out-of-bag evaluation. Full preprocessing pipeline applied independently within each resampling iteration.

3.8. Feature Importance and Explainability Analysis

The three highest-ranked features — loss of self-control, unconscious time loss, and mental preoccupation — show the strongest predictive association with self-reported disengagement difficulty in this dataset. These features conceptually correspond to the salience and withdrawal components of Griffiths' model [4]. However, these associations should not be interpreted as evidence of construct validity for the survey instrument, as the instrument itself has not been independently validated against a clinical reference standard. Nor should the SHAP rankings be interpreted causally: the cross-sectional design of this study precludes any determination of whether these features cause disengagement difficulty or are simply correlated with it within this specific sample. The SHAP findings are best understood as identifying which features most strongly co-vary with the target label in this dataset, under the specific classifiers and feature encoding choices used.

Table 9. Feature Importance Ranking (Mean |SHAP| across supervised models, out-of-fold predictions)

No	Rank	Feature (Indonesian)	Description	Mean SHAP	RF MDI (normalized)
1	1	Tidak_Bisa_Kontrol_Diri	Inability to control use	0.142	0.183
2	2	Tidak_Sadar_Waktu	Unconscious time loss	0.128	0.161
3	3	Terpikirkan	Mental preoccupation	0.114	0.148
4	4	Tanpa_Tujuan	Purposeless browsing	0.097	0.122
5	5	Waktu_Harian	Daily usage duration	0.089	0.103
6	6	Fomo	Fear of missing out	0.076	0.091
7	7	Membuang_Waktu	Perceived time wasting	0.063	0.078
8	8	Sering_Instagram	Instagram as primary platform	0.051	0.044

Note: SHAP values are mean absolute values computed via TreeExplainer (RF, XGBoost) and KernelExplainer (SVM, LR) on out-of-fold predictions. KernelExplainer values for SVM and LR are approximate. MDI normalized to sum to 1.0 within RF. SHAP values represent predictive associations, not causal relationships.

3.9. Discussion

SVM achieved the highest performance across all primary metrics under the tested configuration, dataset, and hyperparameters. However, as established in Section 4.4, this advantage is not statistically significant over Random Forest (corrected t $p=0.102$; McNemar $p=0.073$) and should not be stated as a definitive finding. The competitive performance of Logistic Regression — a simpler linear classifier — alongside SVM further suggests that the classification task in this feature space may favor linear decision boundaries, though this hypothesis requires verification through dimensionality reduction visualization (e.g., PCA) and comparison with kernel SVM variants before it can be stated with confidence. At this sample size, the statistical tests lack the power to

distinguish between SVM, Logistic Regression, and Random Forest. A substantially larger dataset would be required to resolve these comparisons.

SVM's weighted recall standard deviation of 0.244 (Var=0.060) is notably the highest among all evaluated models and substantially exceeds Random Forest (SD=0.183) and XGBoost (SD=0.148). This high variability is a material concern. It indicates that SVM's sensitivity to the positive class fluctuates considerably across different training set compositions — in some folds performing strongly and in others poorly — which implies that its aggregate recall figure masks substantial unpredictability. In any hypothetical future application where consistent identification of students reporting disengagement difficulty was desired, this instability would be a practical disqualifier for SVM as the sole recommended model without further validation. Random Forest's combination of competitive recall (71.6%) with substantially lower variance (SD=0.183) makes it a more practically stable choice under the tested conditions.

XGBoost's lower performance (accuracy 66.8%, MCC=0.281) is consistent with the hypothesis that gradient boosting algorithms require larger training sets to accumulate statistically stable iterative corrections. With approximately 106 training instances per fold, the 100-iteration boosting sequence may overfit to fold-specific noise, which is reflected in XGBoost's higher accuracy SD (0.155). The low-dimensional feature space (approximately 20 features after encoding) may also reduce the benefit of column subsampling. These are plausible mechanistic explanations, but they remain hypotheses: confirming them would require systematic ablation experiments varying sample size and feature dimensionality, and applying optimized hyperparameter configurations via grid search. XGBoost's underperformance on this dataset does not constitute evidence of general inferiority on tabular data.

The SHAP analysis identifies loss of self-control, unconscious time loss, and mental preoccupation as features most strongly associated with the target label. These features are conceptually aligned with salience and withdrawal dimensions in behavioral addiction frameworks, which provides a degree of face validity. However, the following limitations apply to this interpretation. First, the features are drawn from an adapted — not validated — instrument, meaning the face validity claim is preliminary. Second, SHAP values reflect predictive associations within this specific model, dataset, and encoding; they do not

establish that these features are causal determinants of disengagement difficulty. Third, the SHAP computation for SVM and Logistic Regression used KernelExplainer, which produces approximate values; the rankings for these models carry additional uncertainty. Fourth, the finding that daily usage duration ranks fifth below self-regulatory items is consistent with the theoretical distinction between heavy use and problematic use, but cannot be treated as strong evidence for this distinction without a validated instrument and larger sample.

Instagram's status as the most frequently reported platform (58.5%) and its low but nonzero SHAP ranking (0.051) suggest a weak predictive association between platform preference and disengagement difficulty. Given the low importance value and the approximate nature of SHAP for non-tree models, this finding should be interpreted with caution. It is plausible that Instagram's design features contribute to the behavioral patterns captured by higher-ranked items, but this hypothesis is not testable from the current data and should be addressed in future studies incorporating platform-specific behavioral metrics.

To maintain consistency with the study's stated scope, any application of these classifiers in institutional counseling or student support contexts must be treated as a hypothetical future scenario, not a present recommendation. The study explicitly does not support deployment at this stage. However, for the benefit of future researchers, it is useful to note the error cost asymmetry: false negatives (failing to identify a student reporting disengagement difficulty) carry higher practical cost than false positives (unnecessarily flagging a student who does not), because the former results in a student receiving no support while the latter results in an unnecessary but relatively low-harm contact. If a future validated system were deployed, recall on the positive class would therefore be the primary optimization criterion. Under the current results, SVM achieves the highest average positive-class recall (0.827) but with the highest inter-fold variability; Random Forest achieves marginally lower recall (0.853 from confusion matrix, noting this is a fold-averaged approximation) with greater consistency. These trade-offs are presented as analytical observations for future research design, not as deployment guidance.

A critical limitation of all performance estimates in this study is the absence of external validation. All reported metrics are internal cross-validation estimates on 118 respondents

recruited through convenience sampling from a limited number of Indonesian institutions. No held-out independent dataset from a different institution, cohort, or time period was used to evaluate generalizability. Cross-validation, even when repeated and bootstrapped, estimates how well the model learns the patterns in the available data – it does not guarantee that those patterns will reproduce in new data collected under different conditions. External validation on an independent sample is the necessary next step before any conclusions about real-world model utility can be drawn.

The convenience sampling approach introduces two related biases. First, the sample may not represent the Indonesian university student population at large: students who are active in digital organization networks and responsive to social media survey invitations may have systematically higher social media engagement than non-respondents, potentially inflating the positive class prevalence (63.6%) relative to the true population rate. Second, if training data prevalence differs from deployment prevalence, classifier decision boundaries optimized on this sample may produce different false positive and false negative rates in deployment. Future studies should apply probability-based sampling designs and report prevalence estimates with appropriate uncertainty bounds.

4. CONCLUSION

This study compared Random Forest, XGBoost, linear Support Vector Machine (SVM), and Logistic Regression for predicting students' self-reported difficulty disengaging from social media using 118 survey responses and 10-fold stratified cross-validation. SVM achieved the highest mean accuracy (77.2%), ROC-AUC (0.804), and MCC (0.435); however, its performance was not significantly different from Random Forest, which showed greater stability across resampling procedures. SVM performed significantly better than XGBoost, although this result should be interpreted cautiously because of the small sample size, class imbalance, default hyperparameter settings, and internal validation design. Feature-importance analyses consistently identified loss of self-control, unrecognized time spent, and mental preoccupation as the strongest predictors; these associations are predictive rather than causal. Importantly, the target variable was a single self-reported behavioral proxy and should not be interpreted as a clinically validated measure of social media addiction. Therefore, the evaluated models are not yet suitable for deployment in counseling, student support, or early-warning systems. Future

research should use validated multi-item psychometric scales, larger and more diverse multi-institutional samples, longitudinal designs, systematic hyperparameter optimization, external validation, and ethical review before considering practical implementation.

REFERENCES

- [1] S. Kemp, "Digital 2024: Indonesia," *DataReportal*, Feb. 2024. [Online]. Available: <https://datareportal.com/reports/digital-2024-indonesia>. Accessed: Aug. 1, 2024.
- [2] S. Kemp, "Digital 2024: Global overview report," *DataReportal*, Jan. 2024. [Online]. Available: <https://datareportal.com/reports/digital-2024-global-overview-report>. Accessed: Aug. 1, 2024.
- [3] C. S. Andreassen and S. Pallesen, "Social network site addiction: An overview," *Curr. Pharm. Des.*, vol. 20, no. 25, pp. 4053–4061, 2014, doi: 10.2174/13816128113199990616.
- [4] M. D. Griffiths, "A 'components' model of addiction within a biopsychosocial framework," *J. Subst. Use*, vol. 10, no. 4, pp. 191–197, 2005, doi: 10.1080/14659890500114359.
- [5] D. J. Kuss and M. D. Griffiths, "Social networking sites and addiction: Ten lessons learned," *Int. J. Environ. Res. Public Health*, vol. 14, no. 3, Art. no. 311, 2017, doi: 10.3390/ijerph14030311.
- [6] C. Marino, G. Gini, A. Vieno, and M. M. Spada, "The associations between problematic Facebook use, psychological distress and well-being among adolescents and young adults: A systematic review and meta-analysis," *J. Affect. Disord.*, vol. 226, pp. 274–281, 2019, doi: 10.1016/j.jad.2017.10.007.
- [7] C. S. Andreassen, T. Torsheim, G. S. Brunborg, and S. Pallesen, "Development of a Facebook addiction scale," *Psychol. Rep.*, vol. 110, no. 2, pp. 501–517, 2012, doi: 10.2466/02.09.18.PR0.110.2.501-517.
- [8] A. J. van Deursen, C. L. Bolle, S. M. Hegner, and P. A. M. Kommers, "Modeling habitual and addictive smartphone behavior: The role of smartphone usage types, emotional intelligence, social stress, self-regulation, age, and gender," *Comput. Hum. Behav.*, vol. 45, pp. 411–420, 2015, doi: 10.1016/j.chb.2014.12.039.
- [9] F. Jiang *et al.*, "Artificial intelligence in healthcare: Past, present and future," *Stroke Vasc. Neurol.*, vol. 2, no. 4, pp. 230–243, 2017, doi: 10.1136/svn-2017-000101.

- [10] D. B. Dwyer, P. Falkai, and N. Koutsouleris, "Machine learning approaches for clinical psychology and psychiatry," *Annu. Rev. Clin. Psychol.*, vol. 14, pp. 91–118, 2018, doi: 10.1146/annurev-clinpsy-032816-045037.
- [11] D. Loureiro and M. Veloso, "Machine learning for social media mental health classification: A systematic review," *IEEE Access*, vol. 10, pp. 14125–14140, 2022, doi: 10.1109/ACCESS.2022.3147714.
- [12] K. Siste *et al.*, "The impact of physical distancing and associated factors towards internet addiction among adults in Indonesia during COVID-19 pandemic: A nationwide web-based study," *Front. Psychiatry*, vol. 11, Art. no. 580977, 2020, doi: 10.3389/fpsy.2020.580977.
- [13] D. Lazer *et al.*, "Computational social science," *Science*, vol. 323, no. 5915, pp. 721–723, 2009, doi: 10.1126/science.1167742.
- [14] G. Harman, N. Zain, and R. Mahadzir, "Social media addiction prediction using machine learning classifiers," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 5, pp. 478–487, 2021, doi: 10.14569/IJACSA.2021.0120555.
- [15] N. Akhter and A. Nayab, "Predicting problematic social media use using ensemble machine learning," *Comput. Hum. Behav. Rep.*, vol. 6, Art. no. 100188, 2022, doi: 10.1016/j.chbr.2022.100188.
- [16] C. S. Andreassen, S. Pallesen, and M. D. Griffiths, "The relationship between addictive use of social media, narcissism, and self-esteem: Findings from a large national survey," *Addict. Behav.*, vol. 64, pp. 287–293, 2017, doi: 10.1016/j.addbeh.2016.03.006.
- [17] R. J. J. M. van den Eijnden, J. S. Lemmens, and P. M. Valkenburg, "The Social Media Disorder Scale," *Comput. Hum. Behav.*, vol. 61, pp. 478–487, 2016, doi: 10.1016/j.chb.2016.03.038.
- [18] V. Agarwal, P. Mani, and S. Bodduluri, "Comparative performance of random forest and decision tree in health classification," *J. Med. Syst.*, vol. 46, no. 3, Art. no. 47, 2022, doi: 10.1007/s10916-022-01820-w.
- [19] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, San Francisco, CA, USA, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.
- [20] H. Gao, L. Wang, and Y. Li, "Predicting internet addiction among university students using gradient boosting and random forest," *Comput. Educ.: Artif. Intell.*, vol. 3, Art. no. 100064, 2022, doi: 10.1016/j.caeai.2022.100064.

- [21] A. B. R. Shatte, D. M. Hutchinson, and S. J. Teague, "Machine learning in mental health: A scoping review of methods and applications," *Psychol. Med.*, vol. 49, no. 9, pp. 1426–1448, 2019, doi: 10.1017/S0033291719000151.
- [22] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
- [23] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995, doi: 10.1007/BF00994018.
- [24] J. C. Nunnally and I. H. Bernstein, *Psychometric Theory*, 3rd ed. New York, NY, USA: McGraw-Hill, 1994.
- [25] World Medical Association, "World Medical Association Declaration of Helsinki: Ethical principles for medical research involving human subjects," *JAMA*, vol. 310, no. 20, pp. 2191–2194, 2013, doi: 10.1001/jama.2013.281053.