Bulletin of Informatics and Data Science
Vol. 5 No. 1, May 2026, Page 54−66
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v5i1.153
https://ejurnal.pdsi.or.id/index.php/bids/index

Interpretable Decision Tree for Predicting High Self-Reported Happiness Using Digital Behavior and Lifestyle Indicators

Rafika Damayanti Sururin Nufus[1,*], Fiandra Lazuart Adi Hafizh Arraszaq[2]

[1] Data Science, Telkom University, Surabaya, Indonesia.
[2] Informatics, Telkom University, Surabaya, Indonesia.
Email: [1,*]rafikadamayanti@telkomuniversity.ac.id, [2]fiandralazuartadi@student.telkomuniversity.ac.id
Correspondence Author Email: rafikadamayanti@telkomuniversity.ac.id

ARTICLE INFORMATION

ARTICLE HISTORY:
Submited : March 27, 2026
Revised : May 27, 2026
Accept : May 30, 2026
Publish : May 30, 2026

KEYWORD

Digital Well-Being;
Happiness Classification;
Interpretable Machine Learning;
Decision Tree;
Nested Cross-Validation;
Calibration

CORRESPONDENCE AUTHOR

Email:
rafikadamayanti@telkomuniversity.ac.id

A B S T R A C T

Research on digital well-being is dominated by text-derived features and post hoc explanations, leaving limited evidence on whether compact behavioral data can support an inherently interpretable happiness classifier without obscuring predictive trade-offs. This study develops and audits an inherently interpretable decision tree for classifying high self-reported happiness from age, screen time, sleep quality, stress, offline days, exercise, gender, and social-media platform. The contribution is a leakage-safe evaluation that combines nested cross-validation, comparison with Random Forest, logistic regression, and radial-basis-function support vector machine, out-of-fold discrimination and calibration, a one-standard-error tree-selection rule, bootstrap stability, and sensitivity analyses. Across five outer folds, the decision tree obtained balanced accuracy 0.779 ± 0.049, macro-F1 0.778 ± 0.050, ROC-AUC 0.869 ± 0.045, and Brier score 0.146 ± 0.028. Random Forest achieved the highest mean balanced accuracy (0.820), whereas logistic regression achieved the highest ROC-AUC (0.913) and lowest Brier score (0.122). The selected depth-three tree retained eight leaves and used daily screen time and stress as its primary decision pathways. In 200 bootstrap samples, stress was selected in 100.0% of trees and screen time in 86.5%; however, the root alternated between stress (53.5%) and screen time (45.0%), indicating stable relevance but unstable ordering. Results remained similar with the four-feature subset, while performance changed across happiness thresholds. The model therefore offers transparent decision rules at a modest predictive cost, but the cross-sectional, single-dataset design supports predictive association rather than causal or clinical interpretation.

1. INTRODUCTION

Digital platforms generate behavioral traces that can support prediction, decision making, and theory development in computational social science [1]. In digital well-being research, social-media use has been associated with happiness, psychological distress, life satisfaction, and related outcomes through self-reported use, linguistic content, interaction patterns, and platform behavior [2]. However, the strength and direction of these associations depend on how exposure and well-being are measured, which population is studied, and whether the analysis is designed for explanation or prediction. Recent evidence therefore cautions against treating cross-sectional associations as universal or causal effects [3]. This distinction is important for compact behavioral datasets because screen time, sleep quality, stress, exercise, and offline behavior may represent overlapping indicators of broader personal and contextual conditions. A useful predictive study should consequently evaluate not only whether a model classifies well, but also whether its decision process is transparent, whether its probability estimates are reliable, and whether its interpretation remains stable under reasonable analytical changes.

Recent studies demonstrate the value of machine learning for mental-health and well-being prediction, while also revealing limitations that motivate the present research. Liu et al. reviewed machine-learning approaches for detecting and measuring depression from social-media data [4]. Their review confirmed that online data can support prediction, but the reviewed literature was primarily concerned with depression and relied heavily on text or platform-generated signals. It therefore offers limited evidence on whether a small set of directly interpretable behavioral and lifestyle indicators can classify a positive well-being outcome such as happiness. Han et al. developed a model for predicting psychological well-being from social-media language [5]. Their study showed that linguistic features contain useful information, but language representations are high-dimensional, context-dependent, and difficult to translate into short decision rules that nontechnical readers can inspect. These two studies establish the predictive potential of digital traces, but leave unresolved whether compact structured variables can provide an understandable alternative to text-based modeling.

A second group of studies highlights limitations related to platform dependence and the form of explainability. Hinduja et al. proposed a proactive social-sensor service for mental-health monitoring using Twitter data [6]. The system illustrates the value of real-time social sensing, but its reliance on a specific platform and continuous text streams limits applicability when only cross-platform preferences and self-reported behavioral indicators are available. Mimikou et al.

Bulletin of Informatics and Data Science
Vol. 5 No. 1, May 2026, Page 54−66
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v5i1.153
https://ejurnal.pdsi.or.id/index.php/bids/index

applied explainable machine learning to depression prediction and demonstrated how feature-attribution methods can clarify the outputs of a fitted model [7]. Such post hoc explanations are useful, but they explain a model after training rather than making the prediction mechanism itself directly readable. Reviews of explainable artificial intelligence in healthcare distinguish this approach from intrinsic interpretability, in which the model structure can be inspected without a separate explanation layer [8]. For digital well-being research, this distinction matters because transparent thresholds and decision paths may be easier to audit than feature-attribution values produced by a more complex predictor.

Decision-tree research provides a closer methodological foundation, but important gaps remain. Battista et al. used decision trees to identify parsimonious subgroups and interactions in youth mental-health survey data [9]. Their findings show that recursive partitioning can summarize complex relationships in an accessible form. Nevertheless, their study focused on population-health surveillance rather than digital well-being or high-happiness classification, and it did not combine the interpretable model with probability calibration, bootstrap stability, or sensitivity to alternative definitions of the outcome. Interpretation also requires caution because different feature-importance methods can emphasize different predictors [10]. In addition, comparative research on mental-illness detection shows that model families involve trade-offs between predictive performance, complexity, and interpretability [11]. These limitations make a common evaluation design essential. Nested cross-validation can separate hyperparameter selection from outer-fold performance estimation, reducing the risk that tuning information produces overly optimistic results [12]. Thus, transparency should be evaluated together with fair benchmarking, probability quality, and resampling stability rather than treated as sufficient evidence on its own.

The reviewed literature reveals four connected research gaps. First, existing social-media mental-health studies are dominated by linguistic, textual, or platform-specific data, whereas compact structured indicators of digital behavior and lifestyle remain less frequently evaluated for positive well-being classification [4]–[6]. Second, many explainable applications rely on post hoc feature attribution, leaving limited evidence on whether an interpretable-by-design model can provide useful performance while keeping the complete decision process visible [7]–[9]. Third, transparent models are not always compared with stronger ensemble, linear, and kernel alternatives under identical preprocessing, tuning, and validation conditions; consequently, the predictive cost of interpretability is often unclear [9], [11], [12]. Fourth, a single fitted explanation may appear convincing even when the selected variables, root split, or outcome definition changes across samples and analytical choices [10], [12]. These gaps indicate the need for an integrated framework that connects structured behavioral prediction, intrinsic interpretability, fair model comparison, calibration, and stability assessment in one analysis.

Accordingly, this study develops and audits an inherently interpretable Decision Tree for classifying high self-reported happiness from age, gender, daily screen time, sleep quality, stress level, offline days, exercise frequency, and preferred social-media platform. The empirical contribution is to determine whether these compact variables contain sufficient predictive information without relying on social-media text. The methodological contribution is an end-to-end evaluation of the transparency-performance trade-off. The Decision Tree is compared with Random Forest, logistic regression, and radial-basis-function support vector machine under the same leakage-safe nested cross-validation design. Evaluation includes threshold-based classification, ROC-AUC discrimination, Brier-score probability error, calibration, and out-of-fold confusion patterns. Bootstrap analysis examines the stability of feature use and root selection, while sensitivity analyses evaluate alternative happiness thresholds and a reduced predictor set. The objective is not to claim that a single tree is universally superior, but to quantify how much predictive performance is exchanged for readable rules and whether those rules remain reasonably stable. Because the data are cross-sectional and self-reported, the resulting thresholds are interpreted as predictive decision boundaries rather than causal, clinical, or prescriptive cutoffs.

2. RESEARCH METHODOLOGY

2.1 Research Design and Workflow

The analysis was implemented in Python using a reproducible notebook and scikit-learn pipelines [13]. Figure 1 summarizes the sequence from data audit and target construction through nested model comparison, one-standard-error tree selection, out-of-fold evaluation, bootstrap stability analysis, and sensitivity analysis. All preprocessing and hyperparameter selection occurred inside resampling folds so that validation observations could not influence model development [14].

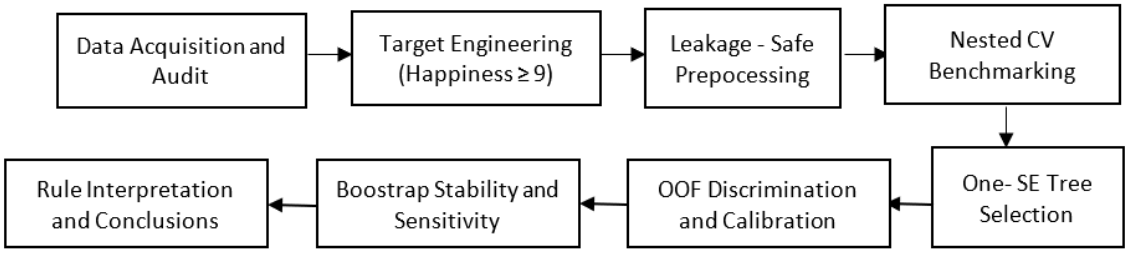

Figure 1. Research workflow for model development, evaluation, and interpretation

Bulletin of Informatics and Data Science
Vol. 5 No. 1, May 2026, Page 54−66
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v5i1.153
https://ejurnal.pdsi.or.id/index.php/bids/index

Figure 1 shows that model development and performance estimation are deliberately separated: hyperparameters are chosen only in the inner folds, whereas each outer fold remains untouched until final validation. This separation is important for a dataset of 500 records because preprocessing or tuning outside the fold structure would allow validation information to influence model selection and could produce optimistic performance estimates [14]. After the nested comparison, the workflow evaluates out-of-fold discrimination and calibration, refits a parsimonious tree, and then examines bootstrap stability and analytical sensitivity. Consequently, the study interprets the tree only after testing whether its performance, variable use, and conclusions remain reasonably consistent across resamples and alternative modeling choices [14].

2.2 Dataset, Predictors, and Target Engineering

The Mental Health and Social Media Balance Dataset contains 500 records and nine substantive attributes: age, gender, daily screen time, sleep quality, stress level, days without social media, exercise frequency, preferred social-media platform, and Happiness Index [15]. The audit found no duplicate user identifiers or missing values. Because the variables are self-reported and the sampling frame is undocumented, the analysis treats the data as a methodological classification case study rather than a population estimate or clinical screening instrument. Table 1 defines the analytical role and fold-specific processing of every variable.

Table 1. Dataset variables and analytical roles

Variable	Role and scale	Processing
Happiness Index	Outcome, integer 1-10	Binarized at primary threshold 9; thresholds 8 and 10 tested
Age	Numeric predictor	Passed without scaling for tree; standardized for LR and SVM
Daily screen time	Numeric predictor, hours/day	Same fold-specific processing as age
Sleep quality	Numeric predictor, 1-10	Same fold-specific processing as age
Stress level	Numeric predictor, 1-10	Same fold-specific processing as age
Offline days	Numeric predictor	Same fold-specific processing as age
Exercise frequency	Numeric predictor, sessions/week	Same fold-specific processing as age
Gender	Categorical predictor	One-hot encoded inside each training fold
Social-media platform	Categorical predictor	One-hot encoded inside each training fold

Table 1 shows that each variable has a predefined analytical role rather than being processed uniformly. Numeric predictors enter the tree in their original units so that thresholds remain directly interpretable, while logistic regression and RBF SVM receive fold-specific standardization because their fitted coefficients and distances are scale-sensitive. Gender and platform are one-hot encoded within the current training fold, the user identifier is excluded, and the original Happiness Index is used only to construct the target. This design makes the four algorithms comparable while preventing information from the validation folds from entering the preprocessing stage [14].

For observation i, high self-reported happiness at threshold τ was defined by (1). The primary threshold τ=9 produced 256 High observations (51.2%) and 244 Lower observations (48.8%).

$$\begin{gathered} y_i^{(\tau)} = 1 \text{ if } H_i \geq \tau, \quad \tau = 9 \\ y_i^{(\tau)} = 0 \text{ if } H_i < \tau \\ H_i = \text{Happiness Index for record } i \\ y_i^{(\tau)} = 1\text{: High}; \quad y_i^{(\tau)} = 0\text{: Lower} \end{gathered} \tag{1}$$

The labels High and Lower are operational score groups and are not clinical diagnoses.

2.3 Preprocessing and Categorical Encoding

Numeric variables were passed directly to tree-based estimators. Logistic regression and RBF SVM used numerical standardization fitted only on the current training fold. Gender and platform were transformed by the one-hot encoding in (2), also fitted inside each training fold.

$$\begin{gathered} x_{ik}^{(\text{cat})} = 1 \text{ if } C_i = c_k \\ x_{ik}^{(\text{cat})} = 0 \text{ if } C_i \neq c_k \\ k = 1, \ldots, K; \quad C_i = \text{category for record } i \\ c_k = \text{category } k; \quad K = \text{number of categories} \end{gathered} \tag{2}$$

Bulletin of Informatics and Data Science
Vol. 5 No. 1, May 2026, Page 54−66
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v5i1.153
https://ejurnal.pdsi.or.id/index.php/bids/index

Unknown categories were ignored during transformation, and the identifier plus the original Happiness Index were excluded from the predictor matrix.

2.4 Models, Nested Cross-Validation, and Hyperparameter Search

Four classifiers were benchmarked: Decision Tree, Random Forest, logistic regression, and RBF SVM. The outer evaluation used five stratified folds, while four-fold stratified inner cross-validation selected hyperparameters by macro-F1. Nested resampling estimates the performance of the complete tuning procedure without allowing the outer validation fold to influence model selection [14]. The same outer splits were used for all algorithms, and out-of-fold probabilities and labels were concatenated for aggregate ROC, calibration, and confusion-matrix analyses. Table 2 lists the preprocessing and inner-loop search space applied to each classifier.

Table 2. Benchmark models and inner-loop search spaces

Model	Preprocessing	Hyperparameter grid
Decision Tree	Numeric pass-through; one-hot categorical	criterion={gini, entropy}; depth=2-5; min leaf={5,10,20}; min split={10,20}
Random Forest	Numeric pass-through; one-hot categorical	100 trees; depth={5,None}; min leaf={2,5,10}; max features=sqrt
Logistic Regression	Standardized numeric; one-hot categorical	C={0.1,1,10}
RBF SVM	Standardized numeric; one-hot categorical	C={0.3,1,3}; gamma={scale,0.03,0.1}

Table 2 demonstrates that each comparator was evaluated with preprocessing and tuning appropriate to its mathematical structure. The Decision Tree and Random Forest can learn nonlinear threshold interactions directly, logistic regression estimates an additive linear decision boundary on the log-odds scale, and the RBF SVM can construct a flexible nonlinear boundary after standardization. The comparison therefore tests whether the tree's intrinsic transparency is accompanied by a measurable predictive cost relative to stronger ensemble, linear, and kernel alternatives. Because all algorithms use the same outer folds and the same inner macro-F1 selection criterion, the reported differences are less likely to be artifacts of favorable data partitions [12].

2.5 Decision Tree Theory and Parsimonious Selection

A classification tree recursively searches for a feature j and threshold t that maximize the reduction in class impurity [9]. Entropy and information gain are defined in (3) and (4).

$$\mathcal{H}(S) = -\sum_{c=0}^{1} p_c \log_2 p_c$$
$$p_c = \frac{n_c}{|S|}, \quad c \in \{0,1\} \tag{3}$$
$$\mathcal{H}(S) = 0 \text{ for a pure node}$$

$$\mathrm{IG}(S,j,t) = \mathcal{H}(S) - \frac{|S_L|}{|S|}\mathcal{H}(S_L) - \frac{|S_R|}{|S|}\mathcal{H}(S_R)$$
$$S_L = \{i \in S : x_{ij} \le t\} \tag{4}$$
$$S_R = \{i \in S : x_{ij} > t\}$$
$$j = \text{candidate feature}; \quad t = \text{candidate threshold}$$

The selected split maximizes (4) subject to the depth and minimum-sample constraints, while terminal-node probability and classification are defined in (5).

$$\hat{p}_\ell = \frac{n_{\ell,1}}{n_\ell}$$
$$\hat{y}_\ell = 1 \text{ if } \hat{p}_\ell \ge 0.5; \quad 0 \text{ otherwise} \tag{5}$$
$$n_{\ell,1} = \text{High records in leaf } \ell$$
$$n_\ell = \text{all training records in leaf } \ell$$

The final full-data tree used the one-standard-error rule to avoid unnecessary complexity. Candidate configurations whose cross-validated macro-F1 scores were within one estimated standard error of the best score were considered eligible under the one-standard-error rule, after which the least complex tree was selected. The resulting model used entropy, maximum depth 3, minimum leaf size 5, and minimum split size 20, producing eight terminal leaves.

Bulletin of Informatics and Data Science
Vol. 5 No. 1, May 2026, Page 54−66
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v5i1.153
https://ejurnal.pdsi.or.id/index.php/bids/index

2.6 Evaluation Metrics and Calibration

Threshold-based evaluation used accuracy, sensitivity, specificity, balanced accuracy, class-specific F1, and macro-F1. Probability discrimination used ROC-AUC, while probability error used the Brier score. Reporting discrimination and calibration together is important because models with similar class accuracy can produce probabilities of different reliability [16].

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$
$$TP, TN = \text{correct High and Lower predictions}$$
$$FP, FN = \text{Lower-to-High and High-to-Lower errors}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$
$$\text{Specificity} = \frac{TN}{TN + FP} \tag{7}$$
$$\text{Sensitivity} = \text{recall of observed High records}$$
$$\text{Specificity} = \text{recall of observed Lower records}$$

$$\text{Balanced Accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2} \tag{8}$$
$$\text{Balanced Accuracy} = \text{equal-weight mean of both class recalls}$$

$$F1_c = \frac{2\,\text{Precision}_c\,\text{Recall}_c}{\text{Precision}_c + \text{Recall}_c}$$
$$\text{MacroF1} = \frac{F1_0 + F1_1}{2} \tag{9}$$
$$c \in \{0,1\} \text{ indexes Lower and High classes}$$

$$\text{Brier} = \frac{1}{N}\sum_{i=1}^{N}(\hat{p}_i - y_i)^2 \tag{10}$$
$$\hat{p}_i = \text{out-of-fold probability of High for record } i$$
$$y_i \in \{0,1\}; \quad N = \text{number of evaluated records}$$

Lower Brier scores indicate smaller probability errors, while calibration intercepts near 0 and slopes near 1 indicate closer agreement between predicted and observed probabilities [16].

2.7 Stability and Sensitivity Analyses

Model stability was assessed through 200 bootstrap resamples. Each resample refitted the selected tree pipeline and recorded both the root feature and whether each feature appeared anywhere in the tree. The two frequencies are defined in (11).

$$\hat{\pi}_j^{\text{root}} = \frac{1}{B}\sum_{b=1}^{B} I\,(R_b = j)$$
$$\hat{\pi}_j^{\text{select}} = \frac{1}{B}\sum_{b=1}^{B} I\,(j \in T_b) \tag{11}$$
$$I(A) = 1 \text{ if } A \text{ is true}; \quad I(A) = 0 \text{ otherwise}$$
$$B = 200; \quad R_b = \text{root feature in tree } b$$
$$T_b = \text{features selected in tree } b$$

These frequencies quantify resampling stability rather than causal importance [10]. Sensitivity analysis repeated nested evaluation at thresholds 8, 9, and 10 and compared all predictors with age, screen time, sleep quality, and stress.

3. RESULTS AND DISCUSSION

3.1 Outcome Distribution and Descriptive Associations

At the primary threshold, 256 records were High and 244 were Lower, so the two classes were nearly balanced. Figure 2 summarizes the score distribution and the descriptive relationships of screen time, stress, and sleep quality with the Happiness Index.

Bulletin of Informatics and Data Science
Vol. 5 No. 1, May 2026, Page 54−66
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v5i1.153
https://ejurnal.pdsi.or.id/index.php/bids/index

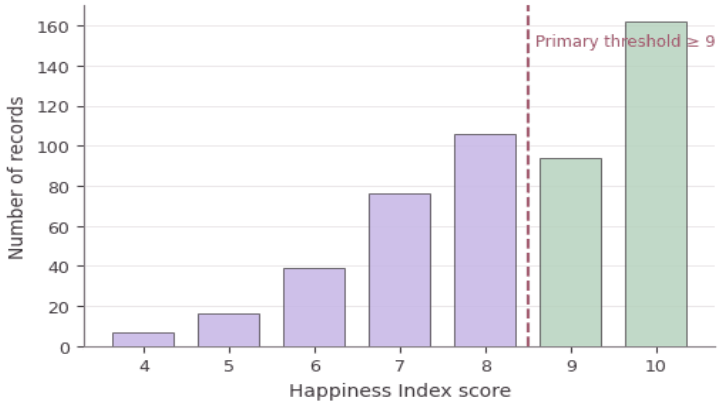

Figure 2. Distribution of Happiness Index scores at the primary threshold

Figure 2 shows that the observed Happiness Index scores are concentrated toward the upper end of the scale. Scores 4–10 contain 7, 16, 39, 77, 105, 94, and 162 records, respectively. Applying the primary threshold of 9 therefore assigns 256 records at scores 9–10 to the High group (51.2%) and 244 records at scores 4–8 to the Lower group (48.8%). These near-balanced split limits majority-class dominance and supports reporting balanced accuracy and macro-F1 in addition to overall accuracy.

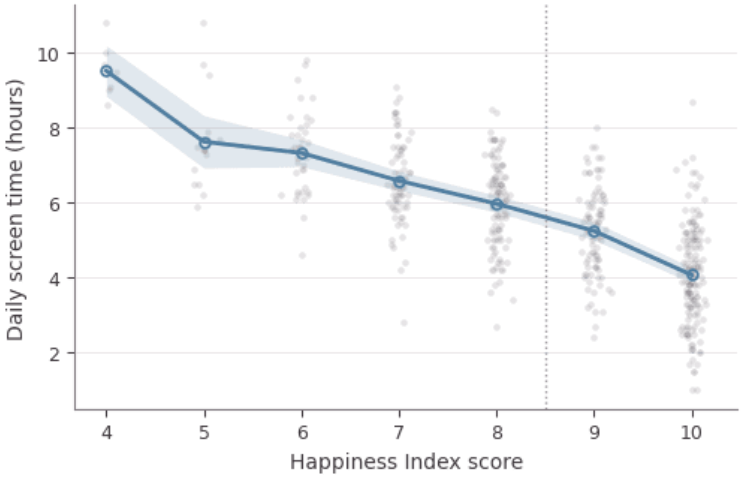

Figure 3. Daily screen time across Happiness Index scores

Figure 3 shows a clear descriptive decline in daily screen time as the Happiness Index increases. Mean screen time falls from 9.53 hours at score 4 to 4.08 hours at score 10, consistent with the negative Pearson correlation of r=-0.71. The visible dispersion within each score nevertheless indicates substantial individual overlap, so screen time alone does not deterministically separate lower and higher happiness observations.

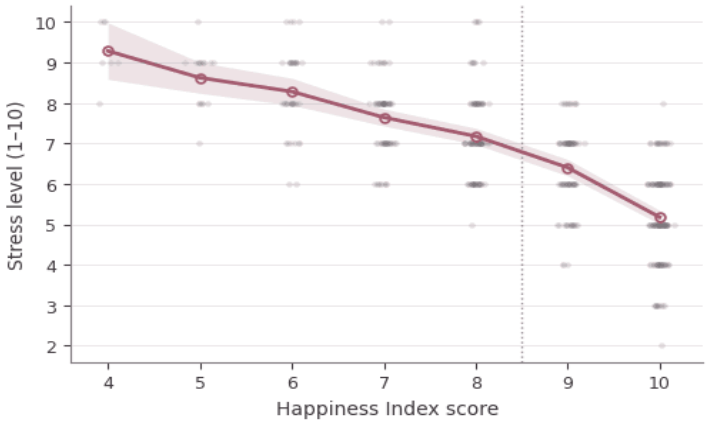

Figure 4. Stress level across Happiness Index scores

Figure 4 shows that mean stress declines from 9.29 at Happiness Index score 4 to 5.18 at score 10. The corresponding correlation of r=-0.74 is the strongest absolute bivariate association with the Happiness Index among the numerical

Bulletin of Informatics and Data Science
Vol. 5 No. 1, May 2026, Page 54−66
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v5i1.153
https://ejurnal.pdsi.or.id/index.php/bids/index

predictors. The horizontal bands reflect the discrete 1–10 stress scale, while the overlap within scores confirms that stress contributes predictive information without acting as a perfect standalone classifier.

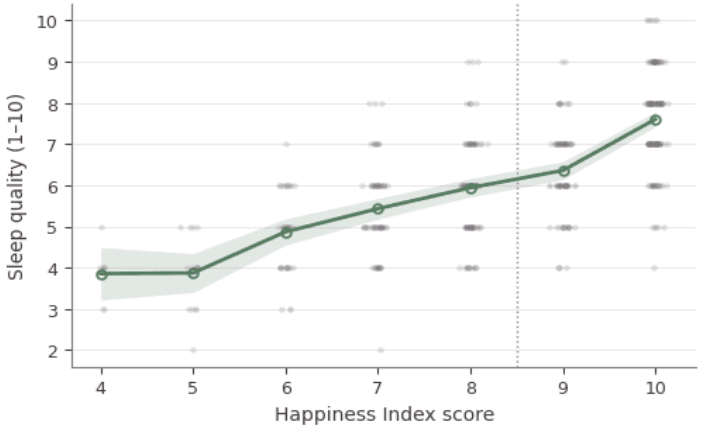

Figure 5. Sleep quality across Happiness Index scores

Figure 5 shows the opposite direction for sleep quality: the mean rises from 3.86 at Happiness Index score 4 to 7.60 at score 10, consistent with r=0.68. Together, Figures 3–5 identify a descriptive profile of lower screen time, lower stress, and better sleep among records with higher self-reported happiness. Because the measures are cross-sectional and self-reported, these patterns are predictive associations rather than evidence that changing any one variable would cause happiness to change [17].

3.2 Comparative Predictive Performance

Table 3 and Figure 6 compare the proposed Decision Tree with Random Forest, logistic regression, and RBF SVM under identical outer folds and inner-loop tuning. The benchmark separates two questions: which model predicts the High/Lower target most effectively, and how much predictive performance is exchanged for an intrinsically readable rule system. Random Forest leads the threshold-based metrics, while logistic regression leads discrimination and probability error. The Decision Tree remains the only benchmark that expresses each prediction as a short threshold-based decision path that can be read directly from a small set of thresholds without a post hoc explanation layer [8].

Table 3. Nested cross-validation performance across five outer folds

Model	Accuracy	Balanced accuracy	Macro-F1	ROC-AUC	Brier ↓
Random Forest	0.820 ± 0.046	0.820 ± 0.047	0.820 ± 0.046	0.899 ± 0.034	0.132 ± 0.020
Logistic Regression	0.816 ± 0.035	0.816 ± 0.035	0.815 ± 0.035	0.913 ± 0.039	0.122 ± 0.023
RBF SVM	0.816 ± 0.032	0.816 ± 0.032	0.815 ± 0.032	0.907 ± 0.043	0.124 ± 0.027
Decision Tree	0.778 ± 0.050	0.779 ± 0.049	0.778 ± 0.050	0.869 ± 0.045	0.146 ± 0.028

Table 3 indicates that Random Forest achieved the highest mean balanced accuracy and macro-F1, both 0.820, exceeding the Decision Tree by 0.041 and 0.042, respectively. Logistic regression and RBF SVM both achieved balanced accuracy of 0.816, while logistic regression produced the highest mean ROC-AUC (0.913) and the lowest Brier score (0.122). The Decision Tree's balanced accuracy of 0.779 and Brier score of 0.146 therefore show a modest but consistent loss in both class-balanced classification and probability quality. Its between-fold standard deviations were also the largest for accuracy and macro-F1 (0.050), indicating greater sensitivity to the composition of the outer training folds. For this study, the tree is justified by transparency rather than predictive superiority, which is consistent with the known variance and accuracy trade-off of single trees relative to ensembles and other flexible classifiers [11].

Figure 6 visualizes the same model trade-off, with horizontal intervals representing between-fold standard deviations rather than population confidence intervals.

Figure 6 makes this trade-off visible: the Decision Tree markers are generally left of the other models for accuracy, balanced accuracy, macro-F1, and ROC-AUC, while its Brier score is farthest to the right, where higher error is worse. The distance is practically modest rather than catastrophic; for example, the tree gives up approximately 4.1 percentage points of balanced accuracy relative to Random Forest while retaining an eight-leaf structure that can be audited rule by rule. The horizontal intervals represent standard deviations across five outer folds, and their overlap means the plot should not be read as a formal significance test or as proof that the ordering will recur in another sample. The value of the benchmark is therefore to quantify the observed transparency-performance exchange under a common resampling design [12].

Bulletin of Informatics and Data Science
Vol. 5 No. 1, May 2026, Page 54−66
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v5i1.153
https://ejurnal.pdsi.or.id/index.php/bids/index

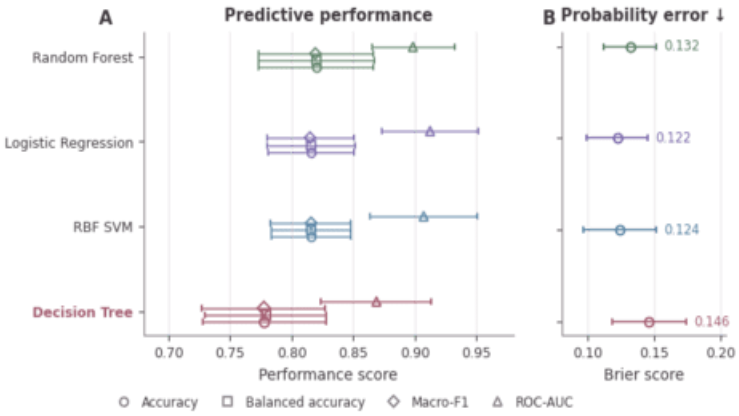

Figure 6. Predictive performance and interpretability trade-off

3.3 Out-of-Fold Discrimination, Calibration, and Confusion Patterns

Figures 7 and 8 separately evaluate aggregate out-of-fold discrimination and probability calibration for the four benchmark models.

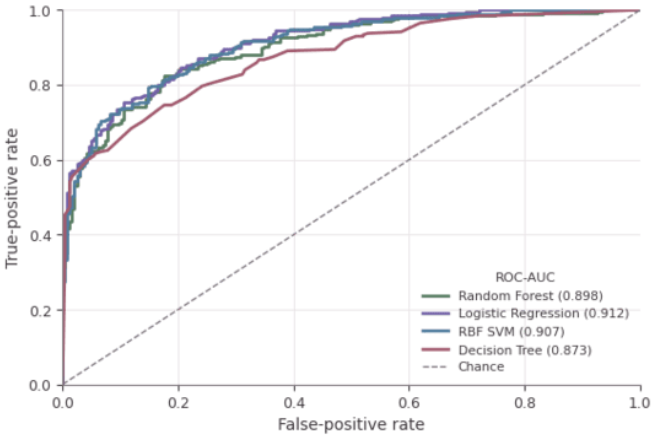

Figure 7. Out-of-fold ROC curves for benchmark classifiers

Figure 7 shows that all four ROC curves lie well above the chance diagonal. Logistic regression provides the strongest aggregate out-of-fold discrimination with ROC-AUC 0.912, followed by RBF SVM at 0.907, Random Forest at 0.898, and Decision Tree at 0.873. The 0.039 difference between logistic regression and the tree indicates that the transparent model loses some ranking ability even before a fixed classification threshold is imposed. In this dataset, an additive logistic model therefore orders the High and Lower records more effectively than the discrete partitions of the selected tree.

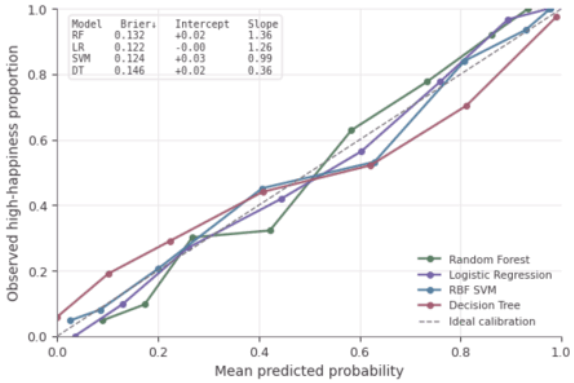

Figure 8. Out-of-fold calibration curves and calibration statistics

Figure 8 shows that logistic regression has the lowest Brier score (0.122), followed by RBF SVM (0.124), Random Forest (0.132), and Decision Tree (0.146). The RBF SVM calibration slope is closest to the ideal value of 1 at 0.99, while its intercept is +0.03. Logistic regression has intercept -0.00 and slope 1.28, Random Forest +0.02 and 1.36, and Decision Tree +0.02 and 0.36. The tree slope indicates that its discrete leaf probabilities are more extreme than the observed out-

Bulletin of Informatics and Data Science
Vol. 5 No. 1, May 2026, Page 54−66
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v5i1.153
https://ejurnal.pdsi.or.id/index.php/bids/index

of-fold frequencies, so its classifications remain interpretable but its raw probabilities should not be treated as calibrated individual risks without external validation or recalibration [16].

3.4 Interpretable Decision Tree and Extracted Rules

The final transparent model was refitted to all 500 records using the one-standard-error configuration, and Figure 9 presents the complete depth-three structure.

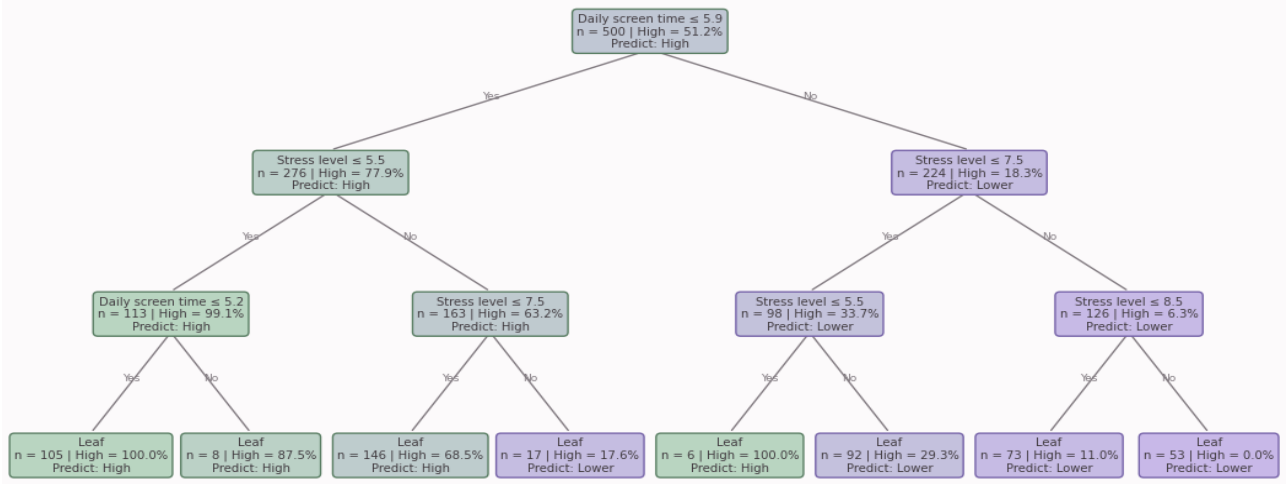

Figure 9. Interpretable decision tree selected by the one-standard-error rule

Figure 9 shows that daily screen time at 5.85 hours is the first partition in the full-data tree. The branch at or below 5.85 hours contains 276 records and has a fitted High proportion of 77.9%, whereas the branch above 5.85 hours contains 224 records and has a fitted High proportion of 18.3%. Stress then refines both branches: among lower-screen-time records, stress at or below 5.5 corresponds to 113 records with 99.1% High, while stress above 7.5 corresponds to 17 records with only 17.6% High. The interaction is especially visible for stress between 5.5 and 7.5, where the fitted High proportion is 68.5% when screen time is at most 5.85 hours but 29.3% when it is higher. In the research context, the tree identifies a joint screen-time and stress profile associated with self-reported happiness, but the thresholds are data-derived decision boundaries rather than causal or clinical cutoffs [9].

Table 4 converts the eight terminal nodes into editable rules with their predicted class, support, and fitted High proportion.

Table 4. Extracted terminal rules from the final decision tree

Rule	Decision path	Predicted class	Support	Fitted High
R2	Screen ≤5.85; Stress ≤5.5; Screen ≤5.25	High	105	1.000
R7	Screen ≤5.85; Stress ≤5.5; Screen >5.25	High	8	0.875
R1	Screen ≤5.85; 5.5<Stress≤7.5	High	146	0.685
R6	Screen ≤5.85; Stress >7.5	Lower	17	0.176
R8	Screen >5.85; Stress ≤5.5	High	6	1.000
R3	Screen >5.85; 5.5<Stress≤7.5	Lower	92	0.293
R4	Screen >5.85; 7.5<Stress≤8.5	Lower	73	0.110
R5	Screen >5.85; Stress >8.5	Lower	53	0.000

Table 4 translates the diagram into eight auditable terminal rules and shows how much data support each rule. R2 contains 105 records with low screen time and low stress and has a fitted High proportion of 100.0%, whereas R5 contains 53 records with screen time above 5.85 hours and stress above 8.5 and has a fitted High proportion of 0.0%. The largest rule, R1, covers 146 records with lower screen time and moderate stress and classifies High with a fitted proportion of 68.5%; by contrast, R3 covers 92 records in the same stress band but with higher screen time and has only 29.3% High. These paired rules make the modeled interaction understandable without a post hoc explainer and show how the tree represents heterogeneous prediction pathways [9]. However, the extreme proportions in R7 (n=8) and R8 (n=6) are based on very small leaves and should not be interpreted with the same confidence as patterns supported by larger leaves. The rules describe associations in the fitted sample and should not be converted into prescriptive screen-time or stress thresholds [8].

Bulletin of Informatics and Data Science
Vol. 5 No. 1, May 2026, Page 54−66
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v5i1.153
https://ejurnal.pdsi.or.id/index.php/bids/index

3.5 Bootstrap Stability of Feature Use

Figure 10 evaluates the reliability of the extracted terminal rules by comparing training-leaf support with class purity.

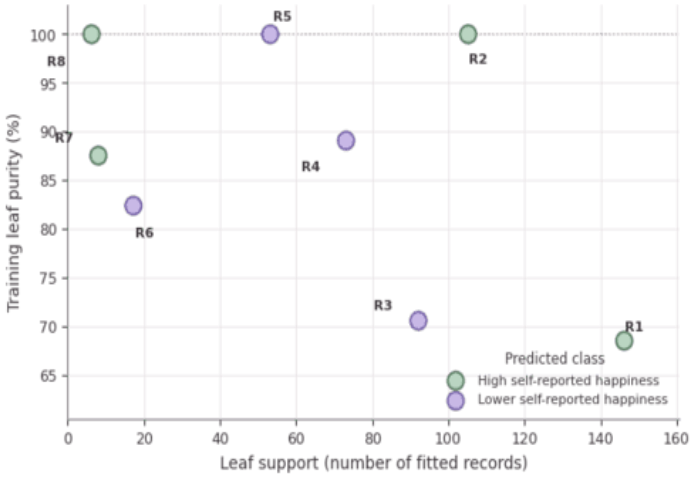

Figure 10. Decision-tree rules by leaf support and training class purity

Figure 10 shows that rule reliability depends on both leaf purity and the number of fitted records supporting the rule. R1 is the largest leaf (n=146) but has 68.5% purity for the High class, whereas R2 (n=105), R5 (n=53), and R8 (n=6) are 100.0% pure; R7 contains only eight records and is 87.5% pure. The remaining Lower-class leaves have 70.7% purity for R3 (n=92), 89.0% for R4 (n=73), and 82.4% for R6 (n=17). Perfect purity therefore does not necessarily imply a dependable rule: R8 is supported by only six observations, while R1 has broad support but substantial class overlap. Separately, the 200 bootstrap refits selected stress in 100.0%, screen time in 86.5%, and sleep quality in 78.0% of trees, while the root alternated between stress (53.5%) and screen time (45.0%), supporting stable joint relevance rather than a uniquely dominant first split [10].

Table 5 details the terminal rules generated by the final Decision Tree and shows how daily screen time and stress jointly determine the predicted happiness class. R2 represents the clearest High-happiness pathway, with stress at or below 5.5 and screen time at or below 5.25 hours, covering 105 records with 100.0% purity. In the moderate-stress range, the comparison between R1 and R3 reveals the main interaction: records are predicted as High when screen time is at or below 5.85 hours, but as Lower when screen time exceeds 5.85 hours. Higher stress combined with longer screen time produces increasingly homogeneous Lower-class rules, reaching 89.0% purity in R4 and 100.0% purity in R5. R6 further indicates that very high stress can lead to a Lower prediction even when screen time remains relatively low. Although R7 and R8 show high purity, their small support of eight and six records requires cautious interpretation. Overall, the rules identify a joint screen-time and stress profile associated with self-reported happiness, rather than independent or causal effects [10].

Table 5. Terminal-leaf support, purity, and predicted class corresponding to Figure 10.

Rule ID	Decision Rule	n	Purity	Predicted Class
R1	Daily_Screen_Time(hrs) ≤ 5.85 AND Stress_Level(1-10) > 5.50 AND Stress_Level(1-10) ≤ 7.50	146	68.5%	High
R2	Daily_Screen_Time(hrs) ≤ 5.85 AND Stress_Level(1-10) ≤ 5.50 AND Daily_Screen_Time(hrs) ≤ 5.25	105	100.0%	High
R3	Daily_Screen_Time(hrs) > 5.85 AND Stress_Level(1-10) ≤ 7.50 AND Stress_Level(1-10) > 5.50	92	70.7%	Lower
R4	Daily_Screen_Time(hrs) > 5.85 AND Stress_Level(1-10) > 7.50 AND Stress_Level(1-10) ≤ 8.50	73	89.0%	Lower
R5	Daily_Screen_Time(hrs) > 5.85 AND Stress_Level(1-10) > 7.50 AND Stress_Level(1-10) > 8.50	53	100.0%	Lower
R6	Daily_Screen_Time(hrs) ≤ 5.85 AND Stress_Level(1-10) > 5.50 AND Stress_Level(1-10) > 7.50	17	82.4%	Lower
R7	Daily_Screen_Time(hrs) ≤ 5.85 AND Stress_Level(1-10) ≤ 5.50 AND Daily_Screen_Time(hrs) > 5.25	8	87.5%	High
R8	Daily_Screen_Time(hrs) > 5.85 AND Stress_Level(1-10) ≤ 7.50 AND Stress_Level(1-10) ≤ 5.50	6	100.0%	High

Bulletin of Informatics and Data Science
Vol. 5 No. 1, May 2026, Page 54−66
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v5i1.153
https://ejurnal.pdsi.or.id/index.php/bids/index

3.6 Sensitivity to Outcome Threshold and Predictor Set

Figures 11 and 12 separately examine sensitivity to the operational outcome threshold and to the predictor set.

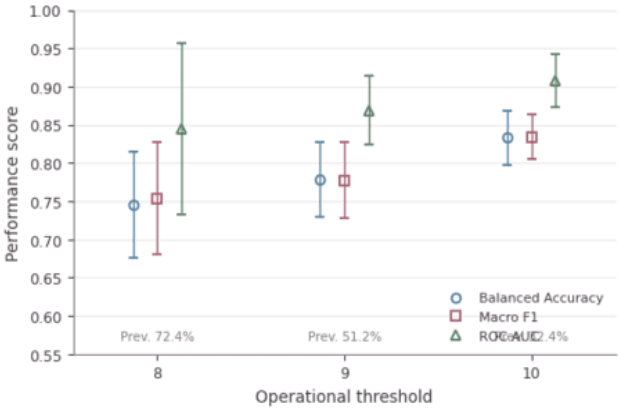

Figure 11. Decision Tree performance across operational outcome thresholds

Figure 11 shows that changing the operational definition of High happiness changes both class prevalence and measured performance. At thresholds 8, 9, and 10, the High class represents 72.4%, 51.2%, and 32.4% of the sample, while balanced accuracy rises from 0.745 to 0.779 and 0.833 and ROC-AUC rises from 0.845 to 0.869 and 0.908. The threshold-10 task therefore appears easier for the tree despite having the smallest positive class, most likely because it contrasts the most extreme happiness scores with the remainder. Each threshold defines a different prediction problem, so the higher values at threshold 10 do not establish that this definition is intrinsically preferable. The result confirms that conclusions depend on how digital well-being is operationalized [17].

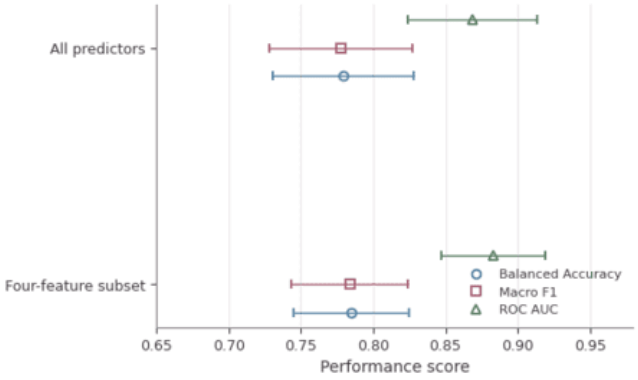

Figure 12. Decision Tree performance using all predictors and a four-feature subset

Figure 12 shows that restricting the model to age, screen time, sleep quality, and stress does not reduce performance. The four-feature subset achieves balanced accuracy 0.785, macro-F1 0.784, ROC-AUC 0.883, and Brier score 0.142, compared with 0.779, 0.778, 0.869, and 0.146 for all predictors. These small improvements suggest that gender, platform, exercise, and offline days add little stable predictive information after the four retained variables are available and may introduce noise into a small-sample tree. The result supports a parsimonious behavioral and lifestyle profile but does not establish that the excluded variables are unimportant in other populations or measurement designs. Feature-set conclusions therefore remain conditional on the present sample, algorithm, and resampling procedure [14].

3.7 Comparison with Previous Research and Study Limitations

Bibliometric evidence confirms rapid growth in social-media mental-health modeling [18]. Survey-based machine learning has predicted human well-being from structured variables [19]. Explainable frameworks have also been evaluated for activity-based depression classification [20]. Workplace mental-health studies similarly use XAI to expose influential predictors [21]. The present study extends this literature by jointly reporting nested benchmarking, calibration, bootstrap stability, and target-definition sensitivity for an interpretable-by-design classifier. Related methodological work has used Random Forest within propensity-score matching to improve covariate balance before estimating exposure–outcome differences in observational health data [22]. This design differs from the present study, which evaluates predictive classification and therefore does not support causal-effect estimation.

Systematic reviews report heterogeneous links between adolescent social-media use and well-being [23]. Person-specific analyses show that estimated effects can differ markedly between individuals [24]. Extended active-passive models propose that how platforms are used may matter more than exposure alone [25]. Mechanistic reviews likewise emphasize vulnerability, context, and developmental pathways [26]. Accordingly, the tree should not be used to prescribe

screen-time limits or to infer that changing a single variable would cause happiness to change. The dataset is cross-sectional, self-reported, limited to 500 records, and lacks a documented sampling frame. The Happiness Index threshold is researcher-defined, subgroup estimates can be unstable, and calibration was evaluated internally rather than on an external population. Future work should use preregistered sampling, objective device logs, longitudinal outcomes, external validation, and probability recalibration before operational deployment.

4. CONCLUSION

This study demonstrates that a compact, inherently interpretable decision tree can classify high self-reported happiness from social-media behavior and lifestyle variables while exposing its predictive trade-offs. Under five-fold nested cross-validation, the tree achieved balanced accuracy 0.779 ± 0.049, macro-F1 0.778 ± 0.050, ROC-AUC 0.869 ± 0.045, and Brier score 0.146 ± 0.028; Random Forest provided higher threshold-based performance, whereas logistic regression provided the strongest discrimination and lowest probability error. The final depth-three tree used screen time and stress as its principal pathways, and bootstrap analysis showed that both variables were consistently relevant even though their root ordering was not stable. A four-feature model performed similarly to the full predictor set, but outcome-threshold sensitivity confirmed that reported performance depends on how High happiness is defined. The contribution is therefore not a claim that the tree is the most accurate model, but evidence that transparent rules can be obtained at a modest predictive cost when evaluation includes fair benchmarking, calibration, stability, and sensitivity analysis. Because the data are cross-sectional, self-reported, drawn from one small public dataset, and internally validated, the findings remain predictive associations and should not be interpreted as causal, clinical, or population-level conclusions.

REFERENCES

[1] J. M. Hofman *et al.*, "Integrating explanation and prediction in computational social science," *Nature*, vol. 595, no. 7866, pp. 181–188, Jun. 2021, doi: 10.1038/s41586-021-03659-0.
[2] D. A. Parry, J. T. Fisher, H. Mieczkowski, C. J. R. Sewall, and B. I. Davidson, "Social media and well-being: A methodological perspective," *Curr. Opin. Psychol.*, vol. 45, p. 101285, Jun. 2022, doi: 10.1016/J.COPSYC.2021.11.005.
[3] J. A. Hall, "Ten Myths About the Effect of Social Media Use on Well-Being," *J. Med. Internet Res.*, vol. 26, no. 1, p. e59585, Nov. 2024, doi: 10.2196/59585.
[4] D. Liu, X. L. Feng, F. Ahmed, M. Shahid, and J. Guo, "Detecting and Measuring Depression on Social Media Using a Machine Learning Approach: Systematic Review.," *JMIR Ment. Health*, vol. 9, no. 3, p. e27244, Mar. 2022, doi: 10.2196/27244.
[5] N. Han *et al.*, "Sensing Psychological Well-being Using Social Media Language: Prediction Model Development Study," *J. Med. Internet Res.*, vol. 25, no. 1, p. e41823, Jan. 2023, doi: 10.2196/41823.
[6] S. Hinduja, M. Afrin, S. Mistry, and A. Krishna, "Machine learning-based proactive social-sensor service for mental health monitoring using twitter data," *International Journal of Information Management Data Insights*, vol. 2, no. 2, p. 100113, Nov. 2022, doi: 10.1016/J.JJIMEI.2022.100113.
[7] C. Mimikou *et al.*, "Explainable Machine Learning in the Prediction of Depression,", vol. 15, no. 11, p. 1412, Jun. 2025, doi: 10.3390/DIAGNOSTICS15111412.
[8] I. D. Mienye *et al.*, "A survey of explainable artificial intelligence in healthcare: Concepts, applications, and challenges," *Inform. Med. Unlocked*, vol. 51, p. 101587, Jan. 2024, doi: 10.1016/J.IMU.2024.101587.
[9] K. Battista, L. Diao, K. A. Patte, J. A. Dubin, and S. T. Leatherdale, "Examining the use of decision trees in population health surveillance research: an application to youth mental health survey data in the COMPASS study," *Health Promotion and Chronic Disease Prevention in Canada*, vol. 43, no. 2, pp. 73–86, Feb. 2023, doi: 10.24095/HPCDP.43.2.03.
[10] M. Saarela and S. Jauhiainen, "Comparison of feature importance measures as explanations for classification models," *SN Applied Sciences*, vol. 3, no. 2, Art. no. 272, Feb. 2021, doi: 10.1007/S42452-021-04148-9.
[11] Z. Ding *et al.*, "Trade-offs between machine learning and deep learning for mental illness detection on social media," *Sci. Rep.*, vol. 15, no. 1, Art. no. 14497 Apr. 2025, doi: 10.1038/s41598-025-99167-6.
[12] J. Wainer and G. Cawley, "Nested cross-validation when selecting classifiers is overzealous for most practical applications," *Expert Syst. Appl.*, vol. 182, p. 115222, Nov. 2021, doi: 10.1016/J.ESWA.2021.115222.
[13] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, 2011, Accessed: Jun. 27, 2026. [Online]. Available: http://jmlr.org/papers/v12/pedregosa11a.html
[14] S. Bates, T. Hastie, and R. Tibshirani, "Cross-Validation: What Does It Estimate and How Well Does It Do It?," *J. Am. Stat. Assoc.*, vol. 119, no. 546, pp. 1434–1445, 2024, doi: 10.1080/01621459.2023.2197686.
[15] "Mental Health & Social Media Balance Dataset." Accessed: Jun. 27, 2026. [Online]. Available: https://www.kaggle.com/datasets/prince7489/mental-health-and-social-media-balance-dataset
[16] L. Liou *et al.*, "Assessing calibration and bias of a deployed machine learning malnutrition prediction model within a large healthcare system," *npj Digit. Med.*, vol. 7, no. 1, Art. no. 149, Jun. 2024, doi: 10.1038/s41746-024-01141-5.
[17] P. M. Valkenburg, "Social media use and well-being: What we know and what we need to know," *Curr. Opin. Psychol.*, vol. 45, p. 101294, Jun. 2022, doi: 10.1016/J.COPSYC.2021.12.006.
[18] J. Kim, D. Lee, and E. Park, "Machine learning for mental health in social media: Bibliometric study," *J. Med. Internet Res.*, vol. 23, no. 3, p. e24870, Mar. 2021, doi: 10.2196/24870.
[19] E. Oparina *et al.*, "Machine learning in the prediction of human wellbeing," *Sci. Rep.*, vol. 15, no. 1, pp. Art. no. 1632, Jan. 2025, doi: 10.1038/s41598-024-84137-1.
[20] I. Ahmed, A. Brahmacharimayum, R. H. Ali, T. A. Khan, and M. O. Ahmad, "Explainable AI for Depression Detection and Severity Classification From Activity Data: Development and Evaluation Study of an Interpretable Framework.," *JMIR Ment. Health*, vol. 12, no. 1, p. e72038, Sep. 2025, doi: 10.2196/72038.

[21] T. Mokheleli, T. Bokaba, and E. Mbunge, "Explainable Artificial Intelligence for Workplace Mental Health Prediction," *Informatics 2025, Vol. 12, Page 130*, vol. 12, no. 4, p. 130, Nov. 2025, doi: 10.3390/INFORMATICS12040130.
[22] R. D. Sururin Nufus, B. Susetyo, B. Sartono, and Efriwati, "Exploring the Potential Impact of Ginger Consumption on the Duration of COVID-19 Recovery: A Propensity Score Matching using Random Forest," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 1359, no. 1, p. 012139, Jun. 2024, doi: 10.1088/1755-1315/1359/1/012139.
[23] M. Shankleman, L. Hammond, and F. W. Jones, "Adolescent Social Media Use and Well-Being: A Systematic Review and Thematic Meta-synthesis," *Adolescent Research Review*, vol. 6, no. 4, pp. 471–492, Apr. 2021, doi: 10.1007/S40894-021-00154-5.
[24] I. Beyens, J. L. Pouwels, I. I. van Driel, L. Keijsers, and P. M. Valkenburg, "Social Media Use and Adolescents' Well-Being: Developing a Typology of Person-Specific Effect Patterns," *Communic. Res.*, vol. 51, no. 6, pp. 691–716, Aug. 2024, doi: 10.1177/00936502211038196.
[25] P. Verduyn, N. Gugushvili, and E. Kross, "Do Social Networking Sites Influence Well-Being? The Extended Active-Passive Model," *Curr. Dir. Psychol. Sci.*, vol. 31, no. 1, pp. 62–68, Feb. 2022, doi: 10.1177/09637214211053637.
[26] A. Orben, A. Meier, T. Dalgleish, and S. J. Blakemore, "Mechanisms linking social media use to adolescent mental health vulnerability," *Nat. Rev. Psychol.*, vol. 3, no. 6, pp. 407–423, May 2024, doi: 10.1038/s44159-024-00307-y.