

Predicting Suicidal Ideation Among Youth Using Family, Cultural, School, and Digital-Ecology Variables: A Stacked Ensemble Model with Calibration, Explainability, and Subgroup Fairness Assessment

Elpida. Christou¹, Hsin-Yi. Chang^{2*}, Andrés. Benítez³

¹ Department of Counseling Psychology, University of Cyprus, Nicosia, Cyprus

² Department of Cognitive Psychology, National Cheng Kung University, Tainan, Taiwan

³ Department of Social Psychology, National University of Asunción, San Lorenzo, Paraguay

* Corresponding author email address: hsinyi.chang@ncku.edu.tw

E d i t o r	R e v i e w e r s
Sergii Boltivets  Chief Researcher of the Department of Scientific Support of Social Formation of Youth. Mykhailo Drahomanov University, Ukraine sboltivets@ukr.net	Reviewer 1: Fahime Bahonar  Department of counseling, Universty of Isfahan, Isfahan, Iran. Email: Fahime.bahonar@edu.ui.ac.ir Reviewer 2: Mahdi Khanjani  Associate Professor, Department of Psychology, Allameh Tabataba'i University, Tehran, Iran. Email: khanjani_m@atu.ac.ir

1. Round 1

1.1. Reviewer 1

Reviewer:

In the Study Design and Participants section, the manuscript reports that “classes or academic units within each institution were randomly invited to participate,” but it does not sufficiently describe the multistage sampling procedure. The authors should specify the number of institutions selected within each geographic region and educational stratum, the method used to randomly select institutions and classes, whether sampling probabilities were equal, and whether survey weights were required. If some regions or school types were intentionally oversampled, weighted and unweighted descriptive estimates should be compared. Because the model may later be considered for application across Taiwan, the representativeness of the sample is highly relevant.

In the paragraph describing the sample division, the authors state that “the final sample of 4,812 participants was divided before model fitting into a development sample of 3,368 participants, a calibration sample of 722 participants, and an independent holdout test sample of 722 participants.” The rationale for this 70%/15%/15% allocation should be explained,

particularly because the positive outcome was present in only 151 participants in each of the calibration and test samples. The authors should demonstrate that the number of positive cases was sufficient for stable calibration and subgroup estimates. Reporting learning curves or resampling-based uncertainty would help establish whether predictive performance had stabilized at the available sample size.

In the ensemble-construction paragraph, the authors indicate that “a nonnegative, regularized logistic-regression meta-learner” was used. The precise formulation of the meta-learner should be reported, including whether an intercept was included, which penalty was applied, how the penalty parameter was selected, and whether the nonnegative coefficients were constrained to sum to one. The final stacking coefficients should be presented so readers can determine how much each base learner contributed. If one algorithm dominated the ensemble, the practical advantage of stacking over selecting that single learner should be reconsidered.

Response: Revised and uploaded the new document.

1.2. Reviewer 2

Reviewer:

In the Data Collection Tools section, the sentence “The primary outcome was suicidal ideation during the preceding month, assessed using the validated traditional Chinese version of the Suicidal Ideation Attributes Scale” requires additional psychometric documentation. The authors should identify the source and validation study for the traditional Chinese version, describe any translation or cultural-adaptation procedure, report internal consistency and dimensionality in the present sample, and justify the selected clinical threshold. If the cutoff was derived in a population that differs from Taiwanese students aged 15–24 years, the possibility of outcome misclassification should be explicitly addressed. Sensitivity analyses using alternative thresholds and the continuous score should be presented in sufficient detail.

The sentence “An additional set of brief safety-screening questions assessed current intent, planning, access to means, and immediate ability to remain safe” raises an important ethical and methodological issue. The manuscript should explain whether survey responses were anonymous or confidential but identifiable during completion, how imminent-risk responses were linked to a specific participant, who monitored alerts, how quickly intervention occurred, and how many participants triggered the safety protocol. If the analytical dataset was fully deidentified, the authors must clarify the operational separation between the safeguarding system and the research dataset. Reporting the number of safety referrals and completed risk assessments would strengthen the ethical transparency of the study.

The Data Collection Tools section describes a very large number of scales, subscales, individual items, contextual characteristics, and behavioral variables, but the exact predictor count is not reported. The authors should provide a complete supplementary data dictionary containing every candidate predictor, scoring range, coding rule, source, reliability coefficient, missing-data rate, and inclusion status in each model. The statement that “conceptually overlapping items were combined into validated scale scores or domain composites when appropriate” is too vague because post hoc aggregation can materially alter model performance and interpretability. All feature-construction rules should be reproducible and prespecified.

In the Data Analysis section, the manuscript states that remaining missing values were addressed “using multiple imputation or model-compatible imputation procedures.” This wording is methodologically ambiguous. The authors must identify the exact imputation method used for every model, the number of imputations if multiple imputation was applied, the variables included in the imputation model, and how estimates were pooled across imputations. It should also be clarified whether imputation was performed separately within each cross-validation training fold to prevent information leakage. Using different imputation approaches across algorithms may introduce unfair model comparisons unless the preprocessing pipelines are carefully harmonized.

In the Data Analysis section, the authors report that “feature reduction was conducted inside the resampling procedure to prevent information leakage,” but no feature-reduction method is named. The manuscript should specify whether reduction involved correlation filtering, recursive feature elimination, univariate filtering, embedded regularization, principal components, or another method. The threshold for removing correlated or low-information predictors should be stated, and the

number of retained features should be reported for each learner. Because explainability findings depend strongly on which correlated variables remain in the model, the stability of feature selection across resamples should also be examined.

The description of nested cross-validation requires greater precision. The sentence “Hyperparameters were tuned using nested, grouped, stratified cross-validation within the 3,368-participant development sample” should be expanded to report the numbers of outer and inner folds, the number of repeats, the grouping unit, the stratification procedure, the hyperparameter search strategy, and the optimization metric. It is particularly important to clarify how stratification by outcome was reconciled with grouping by school, because some folds may have had imbalanced event rates. The authors should also state whether all algorithms were tuned using the same outer resamples to support paired performance comparisons.

Response: Revised and uploaded the new document.

2. Revised

Editor’s decision: Accepted.

Editor in Chief’s decision: Accepted.