Predicting Student Stress Using Machine Learning Ensemble Models: A Multi-Criteria Comparison and Complementary Explanatory Analysis

Daniel Cristóbal Andrade-Girón , William Joel Marin-Rodriguez * , Marcelo Gumercindo Zuñiga-Rojas , Abrahán Cesar Neri-Ayala , Edgar Tito Susanibar-Ramírez , Miguel Angel Aguilar-Luna-Victoria

Posted Date: 3 June 2026

doi: 10.20944/preprints202606.0311.v1

Keywords: student stress; student mental health; explainable AI; ensemble learning; calibration; nested cross-validation; educational data mining

 doi:10.20944/preprints202606.0311.v1

Article

Predicting Student Stress Using Machine Learning Ensemble Models: A Multi-Criteria Comparison and Complementary Explanatory Analysis

Daniel Cristóbal Andrade-Girón [1], William Joel Marin-Rodriguez [2,*], Marcelo Gumercindo Zuñiga-Rojas [3], Abrahán Cesar Neri-Ayala [4], Edgar Tito Susanibar-Ramírez [1] and Miguel Angel Aguilar-Luna-Victoria [5]

[1] Department of Formal and Natural Sciences, Universidad Nacional José Faustino Sánchez Carrión, Lima 15136, Peru
[2] Department of Engineering Systems, Computer and Electronics, Universidad Nacional José Faustino Sánchez Carrión, Lima 15136, Peru
[3] Department of Social Sciences and Communication, Universidad Nacional José Faustino Sánchez Carrión, Lima 15136, Peru
[4] Department of Administration and Management, Universidad Nacional José Faustino Sánchez Carrión, Lima 15136, Peru
[5] Department of Mathematics and Statistics, Universidad Nacional José Faustino Sánchez Carrión, Lima 15136, Peru
* Correspondence: wmarin@unjfsc.edu.pe; Tel.: +51990455214

Abstract

Student stress is a significant mental health issue in educational settings; therefore, the development of reliable, calibrated, and interpretable predictive models can help classify observed levels of stress. This study analyzed the public Student Stress Factors dataset (N = 1,100; 20 predictors and 1 target variable) using a supervised pipeline designed to minimize information leakage. This pipeline incorporated stratified partitioning, encapsulated preprocessing, and nested cross-validation applied exclusively to the training subsets. A comparative analysis was conducted on nine ensembles and boosting algorithms designed for tabular data: AdaBoost, Gradient Boosting, Random Forest, Extra Trees, Bagging, Voting, Stacking, XGBoost, and LightGBM. The performance of the model was evaluated using several evaluation metrics, including accuracy, balanced accuracy, F1-weighted, **Matthews' correlation coefficient** (MCC), **receiver operating characteristic—area under the curve (**ROC-AUC) weighted, and Brier score. These metrics were supplemented by the nonparametric Friedman test, which was used for model comparison. The findings indicated that the performance was consistently high. Gradient Boosting achieved the highest average performance in nested cross-validation (accuracy = 89.55 ± 3.16%; F1-weighted = 89.54 ± 3.17%; MCC = 0.845 ± 0.047; ROC-AUC weighted = 98.59 ± 0.92%). XGBoost and LightGBM followed closely behind. In the independent holdout set, the final calibrated model exhibited robust performance metrics (accuracy = 0.8818; F1-weighted = 0.8818; MCC = 0.8237; ROC-AUC weighted = 0.9861). These findings substantiate the model's stability and generalization capability. However, the Friedman test did not reveal significant differences between the algorithms ($\chi^2 = 10.953$; $p = 0.204$). Consequently, the selection of a model should take into account not only its predictive performance but also its computational efficiency, calibration, interpretability, and feasibility of implementation. Despite the pipeline's internal stability and satisfactory holdout performance, the public and cross-sectional nature of the data imposes limitations on causal inference and model transferability. Consequently, external and prospective validations are imperative prior to its integration into institutional early warning systems.

Keywords: student stress; student mental health; explainable AI; ensemble learning; calibration; nested cross-validation; educational data mining

 doi:10.20944/preprints202606.0311.v1

1. Introduction

In the contemporary higher education landscape, student stress has been identified as a predominant challenge to mental health and well-being on a global scale [1]. This phenomenon is characterized by elevated rates of anxiety, depression, emotional exhaustion, and psychological distress among college students [2,3]. These issues have a direct impact on academic performance, retention rates, quality of life, and, in extreme cases, the risk of suicidal ideation and behavior [4].

The distribution of stress and psycho-emotional distress is not uniform across countries, institutions, or sociodemographic groups [5]. A multitude of structural factors have been identified as contributing to the concentration of risk among specific student subgroups. These factors include economic insecurity, unequal access to educational resources, competitive academic demands, and exposure to hostile environments, such as bullying, symbolic violence, and discrimination [6]. In this context, various authors have described a "silent crisis" of mental health in higher education. This crisis is characterized by the inadequacy of traditional counseling and psychological support services in terms of coverage, timeliness, and responsiveness.

Epidemiological projections indicate that the prevalence of distress and emotional disorders among young people and emerging adults will continue to rise [7], with mental health issues ranking among the leading causes of disability-adjusted life years [8]. This scenario entails significant indirect economic costs associated with poor academic performance, dropping out of school, and lost future productivity [9]. Within the context of higher education, this phenomenon is accompanied by mounting pressure on student wellness services and the urgent need to develop proactive monitoring and prevention strategies. These strategies should be designed to identify at-risk students in an early stage and provide targeted, cost-effective interventions.

In response to this challenge, numerous institutions have implemented mental health promotion programs, stress management workshops, academic tutoring, and awareness campaigns [10,11]. However, extant evidence suggests that the sustained impact of these interventions is limited by the absence of systematic student stress monitoring systems, the fragmentation of information sources (academic, psychological, social, and lifestyle-related), and the lack of integration between institutional policies and quantitative risk analyses [12]. Consequently, there is an imperative to transition toward more sophisticated monitoring models that integrate multidimensional information and enable the prioritization of support resources in a transparent, scalable, and data-driven manner.

In this context, machine learning (ML) has emerged as a promising methodological tool for analyzing large volumes of academic, psychosocial, and behavioral data, identifying nonlinear patterns and complex dependency structures associated with stress, anxiety, and other mental health issues among college students [13,14]. Recent systematic reviews have synthesized the use of ML techniques—primarily supervised classifiers—to predict anxiety and stress in the university population. These reviews have highlighted the diversity of attributes, methodologies, and metrics employed, as well as the predominance of algorithms such as support vector machine (SVM) and logistic regression [15–17].

Recent literature on predicting student stress and mental health using ML converges on the use of tree-based and gradient-boosting models (Random Forest, Boosting, CatBoost, among others) to classify levels of academic stress and associated symptoms (depression, anxiety) [18], based on surveys regarding study habits, sleep patterns, workload, and, in some cases, physiological data or interactions with digital platforms (e.g., eye tracking and multimodal records in e-learning environments) [19]. These approaches enable potentially nonintrusive monitoring of stress and cognitive performance [20] and, in low- and middle-income country contexts, have been complemented by conventional statistical analyses to strengthen screening and university mental health policies [21].

Notwithstanding these advances, substantial methodological gaps persist. First, a considerable proportion of studies concentrate on individual ML models (e.g., artificial neural networks [22], SVM [23], or decision trees [24]). These models are susceptible to high data heterogeneity, class imbalance,

 doi:10.20944/preprints202606.0311.v1

the presence of noise, and sample size constraints. Consequently, this increases the risk of overfitting and reduces generalizability [25–27]. Second, performance evaluation frequently relies on aggregated global metrics, such as accuracy or the area under the ROC curve [28], with clearly insufficient attention paid to more informative indicators in highly imbalanced contexts (F1-score, Cohen's Kappa, Matthews' correlation coefficient (MCC)) and to the stability of performance by class [29]. Furthermore, there is a paucity of studies that incorporate statistically rigorous comparison schemes between models. For example, repeated cross-validation combined with nonparametric rank tests (Friedman's test) and post hoc analysis (Nemenyi's test) [30] allows for the determination of whether the observed differences in performance are truly significant and not attributable to sampling chance [31]. Finally, explanatory analysis at the trait level is often in its infancy or peripheral, with limited use of post hoc interpretability methods [32]. This makes it difficult to robustly link the identified predictors to the accumulated body of psychoeducational knowledge on student stress. As a result, the acceptability of these models among health professionals, well-being teams, and academic authorities is limited.

Given these limitations, ensemble learning methods have emerged as robust and flexible alternatives for classifying stress states by combining multiple base classifiers using aggregation schemes such as bagging, boosting, and stacking [33]. Random Forest, Extra Trees, Gradient Boosting, and XGBoost [34] are examples of architectures that have been demonstrated to reduce bias, control variance, and increase model stability, even in the presence of incomplete, heterogeneous, or noise-contaminated data. Furthermore, many of these decision tree-based algorithms integrate naturally with explainable artificial intelligence (XAI) tools [35], such as SHapley Additive exPlanations (SHAP) values and local explanation methods, including Local Interpretable Model-agnostic Explanations (LIME) [36]. This integration facilitates the derivation of global and local explanations for predictions, as well as the identification of risk groups and actionable patterns of psychosocial vulnerability [37]. This study is predicated on the recognition of these lacunae and seeks to address the following question: Can ensemble learning models predict student stress levels and provide interpretable explanations for the factors associated with their predictions? To address this question, a systematic comparison of various ensemble algorithms is conducted using a protocol involving nested cross-validation, holdout evaluation, probabilistic calibration, and the Friedman test. Furthermore, performance is examined using metrics sensitive to class balance, and a complementary explanatory analysis is incorporated using SHAP and LIME.

Within this framework, the objective of the study is to rigorously develop, compare, and validate ensemble learning models for predicting student stress among university students, taking into account psychological, physiological, academic, environmental, and social factors. The primary contribution of this study lies in its methodological approach, which deviates from the conventional benchmarking approach on public data. Instead, the study proposes a multi-criteria evaluation framework that integrates predictive performance, stability, statistical significance, probabilistic reliability, computational cost, and interpretability. This integration enables an assessment of the utility of the models based on their classification ability, methodological robustness, and the feasibility of their application in educational contexts.

1. Literature Review

2.1. Machine Learning Applied to Student Stress and Mental Health

In the preceding two decades, research on stress and mental health among students has undergone a transition from predominantly descriptive approaches based on psychometric scales to predictive models supported by ML algorithms, ensembles, and explainability techniques (XAI). This transition is driven by the necessity to identify at-risk students with greater precision, to integrate multiple domains of information—psychological, academic, behavioral, physiological, and contextual—and to address issues characteristic of educational data, such as heterogeneity, collinearity, moderate dimensionality, and potential imbalance among outcome categories.

 doi:10.20944/preprints202606.0311.v1

Preliminary studies concentrated predominantly on individual classifiers and survey data. A. Marouf et al. [38] developed a model to examine the relationship between the Big Five personality traits and perceived stress levels in a sample of computer science students. Their findings indicated that algorithms such as sequential minimal optimization (SMO) (SVM) and k-nearest neighbors (k-NN) achieved accuracies close to 70%, thereby demonstrating the feasibility of classifying low, moderate, and high stress levels based on personality profiles. In a broader context of mental health, Flesia et al. [39] employed generalized regression models and ML to identify elevated levels of perceived stress, underscoring the significance of emotional stability, positive coping, and internal locus of control as pertinent variables.

Beginning in 2020, the literature expanded its scope to encompass a variety of mental health outcomes among adolescents, young adults, and emerging adults. Naghavi et al. [40] proposed a system based on stable feature selection and a stacked ensemble of decision trees, achieving an AUC close to 0.90 for a risk outcome among university students in the Middle East and North Africa (MENA) region and identifying a compact set of psychological and social support items as relevant predictors. Liu et al. [41] trained a neural network to predict the risk of post-traumatic stress. The training was conducted on a sample of 2,067 Chinese adults, and the network achieved a classification accuracy of nearly 90% for the cases. The study's findings underscore the influence of depression, anxiety, age, coping, and self-efficacy on the post-traumatic stress risk.

Concurrently, recent reviews have indicated a mounting prevalence of ML applications in educational and occupational contexts for the purpose of addressing stress, anxiety, and depression. Mittal et al. [15] conducted a review of the applications of ML and deep learning for stress management, emphasizing the competitive performance of SVM, Random Forest, XGBoost, and deep architectures such as the convolutional neural network (CNN) and long short-term memory (LSTM). In the context of Latin America, Daza et al. [42] conducted a review of 29 studies focusing on the prediction of anxiety and stress in college students. The study concluded that SVM and logistic regression are frequently effective techniques; however, evaluation of these techniques typically relies on precision and accuracy as central metrics. Tariq et al. [43] compared various ML algorithms, including logistic regression, SVM, decision trees, Random Forest, Gradient Boosting, and XGBoost, to classify levels of college stress. The study reported higher accuracy for Random Forest and incorporated SHAP to identify factors such as blood pressure, perceived safety, sleep quality, and teacher–student relationships. Pujadas et al. [44] employed a longitudinal design, utilizing Balanced Random Forest and other classifiers to predict adolescent mental health status. The researchers demonstrated the relevance of school climate, emotional distress, sleep, and social network using SHAP.

This line of research corroborates the viability of ML for predicting stress and psychological distress among students. However, this study also underscores persistent limitations, including a predominance of individual classifiers, context-specific samples, evaluation focused on aggregate metrics, and limited integration between predictive performance, model stability, and factor interpretation. These limitations necessitate the adoption of more rigorous and transparent comparative designs.

2.2. Ensemble Models for Predicting Stress and Psychological Well-Being

While early studies concentrated on individual classifiers, recent literature indicates a discernible shift toward ensemble models and optimized architectures. In the domain of student stress and well-being, Anand et al. [33] developed an ensemble classifier for academic stress levels. This classifier was constructed using survey data regarding study habits and academic workload. The researchers applied oversampling and 5-fold cross-validation in their analysis. The ensemble that achieved the highest accuracy was 93.48%, and it had an F1-score of 93.14%. The ROC curves of this ensemble exceeded 0.90 across all three categories. Pereira et al. [45] compared Random Forest, XGBoost, and SVM to identify burnout risk profiles, with accuracies ranging from 93.82% to 97.53%.

 doi:10.20944/preprints202606.0311.v1

The researchers highlighted the role of psychological distress, emotional regulation, sleep quality, pro-inflammatory diet, and physical activity.

The preference for ensemble methods is not limited to the academic context. Younis et al. [46] conducted an analysis of bagging, boosting, and stacking of k-NN, decision trees, Random Forest, and SVM for multimodal emotion recognition, determining that stacking achieved the highest accuracy. Almadhor et al. [47] developed a stacking model for stress detection on the WESAD (Wearable Stress and Affect Detection) dataset, demonstrating superior performance in terms of accuracy, precision, recall, and F1-score in comparison to conventional models. Hadhri et al. [48] developed a soft-voting classifier that integrates logistic regression, k-NN, SVM, decision trees, and Random Forest to detect stress levels from IoT data. This approach yielded an accuracy of approximately 78%, representing a significant advancement in the field.

From a methodological perspective, ensemble models are attractive for educational and psychosocial data because they have the capacity to capture nonlinear interactions among psychological, academic, physiological, and contextual variables. Nevertheless, the assumption of superiority based solely on descriptive differences in accuracy or F1-score is unwarranted. Such disparities may be attributable to preprocessing, the partitioning strategy, hyperparameter tuning, the management of imbalance, the number of folds, and the complexity of the outcome. Therefore, a comparison of ensembles necessitates the implementation of validation protocols that facilitate the discernment of genuine advancements from variations attributable to sampling randomness or experimental design.

2.3. Metrics for Evaluation, Validation, and Probabilistic Calibration

A thorough evaluation of predictive models in student mental health necessitates an assessment that extends beyond the mere accuracy of the models. In problems involving multiple classes or those with the potential for class imbalance, metrics such as balanced accuracy, F1-macro, F1-weighted, Cohen's Kappa, and MCC facilitate a more precise evaluation of performance stability across categories. Similarly, in scenarios where predictions are designed to inform educational or psychoeducational decisions, probabilistic quality assumes significance. A model may demonstrate proficiency in distinguishing between classes; however, it is essential to note that the production of poorly calibrated probabilities can result in an inaccurate institutional interpretation of risk.

Nevertheless, the extant literature on student stress prediction frequently relies on global metrics such as accuracy and, to a lesser extent, receiver operating characteristic—area under the curve (**ROC-AUC)**. A number of studies have reported ad hoc combinations of metrics, including accuracy and F1 in Anand et al. [33]; precision, recall, and F1 in Campanella et al. [49]; and AUCPR in Haghish et al. [50]. However, there is a paucity of studies that have compared families of ensembles using a coherent set of indicators focused on cross-class stability, classification agreement, and applied utility.

Furthermore, performance differences between models are rarely tested using nonparametric statistical tests on results obtained fold by fold. Although reviews such as those by Mittal et al. [15], Daza et al. [42], and Mahajan et al. [51] document the growing use of bagging, boosting, stacking, and voting, studies that integrate robust validation, statistical comparison using Friedman or post hoc tests, independent holdout evaluation, and computational cost analysis remain limited. This oversight can result in the overinterpretation of marginal differences between algorithms, particularly when models demonstrate comparable performance.

2.4. Explainability and Interpretation of Risk Factors Using XAI

The incorporation of explainability techniques is becoming an increasingly important component of predictive models applied to mental health and education. The utilization of tools such as SHAP and LIME facilitates the exploration of the variables that contribute to both global and local predictions. This enhances model transparency, thereby enabling discussions concerning the plausibility of the model, informed by psychoeducational knowledge. Recent studies by Shanto and

 doi:10.20944/preprints202606.0311.v1

Jony [52], Geng et al. [53], Pujadas et al. [44], and Tariq et al. [43] have employed these techniques to identify factors related to stress, academic adjustment, mental well-being, and educational risk.

However, the application of XAI in this domain is characterized by inconsistency. The application of this method is frequently limited to one or two models, with a lack of integration with probabilistic calibration, stability analysis, statistical comparison, or out-of-sample evaluation. Furthermore, the attributions of SHAP and LIME are contingent upon the algorithm, the variable encoding scheme, the multiclass formulation, and the type of feature employed. Consequently, their interpretation should be constrained to predictive evidence and not causal evidence.

Therefore, XAI should be regarded as a tool for auditing and interpreting model behavior, not as a substitute for clinical, psychoeducational, or causal validation. The value of this approach lies in its ability to examine the plausibility of the variables relevant to the algorithm, the consistency of local explanations with global patterns, and the transparency with which the system's decisions can be communicated to well-being teams, teachers, or academic authorities. This distinction is of particular importance when models are considered for early warning systems or institutional decision support.

2.5. Research Gap and Contribution of this Study

A comprehensive review of the extant literature has yielded five key gaps. First, there is a paucity of systematic comparisons of multiple families of ensemble models under a single robust validation protocol. Second, numerous evaluations persist in utilizing aggregate metrics, allocating comparatively less attention to indicators that are sensitive to class balance, agreement, and performance stability. Third, probabilistic calibration frequently garners minimal attention, despite its pertinence in the interpretation of probabilities within decision support systems. Fourth, the distinction between descriptive advantages and inferentially defensible differences is rarely assessed through the implementation of nonparametric statistical tests. Fifth, the application of explainability via SHAP or LIME is frequently implemented in isolation, without integration with performance, calibration, and psychoeducational plausibility.

In response to these gaps, this study proposes a multi-criteria comparison of ensemble models for the multiclass prediction of student stress. The contribution of this study extends beyond the mere application of algorithms to a public dataset, encompassing a comprehensive integration of nested cross-validation, independent holdout evaluation, hyperparameter optimization, statistical comparison via the Friedman test, computational cost analysis, probabilistic calibration, and complementary explainability through SHAP and LIME. This strategy enables the assessment of not only which model attains the highest mean performance, but also the statistical significance of the differences, the reliability of the probabilities, and the interpretability of the identified predictive factors from a psychoeducational perspective.

Consequently, this study is situated at the nexus of predictive performance, statistical rigor, and interpretability. Its incremental contribution lies in providing a more transparent evaluation framework for student stress models, avoiding the reduction of comparisons to a simple ranking of accuracy and promoting a more responsible assessment of the potential utility of ensemble models in educational contexts.

3. Methodology

3.1. Analytical Design

A supervised ML pipeline was implemented to predict the stress_level outcome based on demographic, behavioral, environmental, academic, and psychosocial variables. The design of the system was informed by three overarching objectives: to obtain an unbiased estimate of performance, to systematically prevent information leakage, and to conduct a reproducible comparison across multiple algorithms using an independent internal validation scheme.

 doi:10.20944/preprints202606.0311.v1

3.2. Data Source

The data source utilized in this study is an open-access dataset accessible on Kaggle: (https://www.kaggle.com/datasets/rxnach/student-stress-factors-a-comprehensive-analysis) (accessed October 22, 2025), which concerns students at the University of Dharan, Nepal. The dataset corresponds to a cross-sectional observational study involving 1,100 students and is structured in tabular format. The model encompasses 20 predictor variables and a 1 target variable, which are grouped into five conceptual domains: psychological, physiological, social, environmental, and academic. These variables are designed to capture a comprehensive set of individual and contextual indicators associated with stress, including but not limited to anxiety, depression, self-esteem, a history of mental health issues, sleep quality, blood pressure, social support, peer pressure, housing conditions, academic performance, study load, and concerns about the professional future. The measurement of psychological variables is typically conducted on ordinal scales ranging from 0 to 30, while the coding of most contextual and academic factors is performed using Likert-type scales from 0 to 5. The target variable is the student's stress level (stress_level), which is classified into three ordinal categories: low, medium, and high. Consequently, the dataset is suitable for formulating a supervised multiclass classification task aimed at analyzing and predicting student stress.

The methodological relevance of this database is substantiated by recent studies that have employed datasets of analogous structure for predictive modeling of student stress. Tariq et al. [43] in *Scientific Reports* developed ML models aimed at predicting stress in higher education. These models were based on psychological, physiological, academic, and social variables, and they were designed to be explainable. In a similar vein, Liu and Yu [54] employed a dataset comprising 1,100 observations and 20 variables to construct predictive and interpretable models in *Frontiers in Psychology*. Their study underscored the significance of factors such as blood pressure, social support, and depression. The available evidence supports the use of multidimensional tabular databases to train, validate, and interpret ML models aimed at the early identification of stress levels in college students.

3.3. Definition of the Outcome and Initial Debugging

The target variable was explicitly defined as stress_level. Prior to the modeling stage, the existence of the specified column and the validity of its labels were verified. Despite the absence of missing values in the outcome variable, the pipeline incorporated a preventive rule to exclude records lacking a valid label. This precautionary measure was implemented to mitigate potential bias and ensure the integrity of the supervised learning process, as imputing responses can introduce bias and compromise the validity of the learning algorithm.

The outcome labels were standardized and encoded using LabelEncoder, applied exclusively to the target variable. Given that stress_level comprised three categories (low, medium, and high), the study was formulated as a multiclass classification task. Despite the inherent ordinal structure of these categories, a nominal multiclass formulation was adopted to ensure uniform comparison of the discriminative performance of the ensemble algorithms evaluated under standard supervised classification metrics. Consequently, the results should be interpreted as predictions of observed stress levels and not as strict ordinal modeling of a latent progression of stress.

3.4. Variable Typing and Preprocessing

A methodological decision of particular significance was the explicit coding of categorical variables. While certain predictors were denoted by numerical codes, their interpretation fell within the domain of qualitative or ordinal categories. Consequently, these predictors should not be indiscriminately treated as continuous variables. Consequently, mental_health_history, headache, blood_pressure, sleep_quality, breathing_problem, noise_level, living_conditions, safety, basic_needs, academic_performance, study_load, teacher_student_relationship, future_career_concerns, social_support, peer_pressure, extracurricular_activities, and bullying were

 doi:10.20944/preprints202606.0311.v1

defined ex ante as categorical, protected from automatic conversion to numerical format, and assigned the category type prior to preprocessing.

This decision was made to prevent inconsistent transformations, such as imputing values based on the median or standardizing variables that, given their substantive meaning, should not be treated as continuous. For the remaining predictors, a conservative numerical conversion was applied, triggered only when the proportion of values that could be validly interpreted as numerical exceeded a predefined threshold.

The preprocessing stage was executed through the implementation of a ColumnTransformer. The numerical block included median imputation and standardization using StandardScaler, while the categorical block incorporated mode imputation and one-hot encoding using OneHotEncoder(handle_unknown="ignore"). All preprocessing activities were maintained within the confines of the modeling pipeline, thereby ensuring that imputers, scalers, and encoders were trained exclusively on training data residing within each partition.

3.5. Data Partitioning and Imbalance Control

The dataset was stratified into two subsets: a training set comprising 80% of the observations, which were used for model development, tuning, and comparison, and an independent internal holdout set comprising 20%, which was reserved for final evaluation. The stratification process effectively preserved the relative distribution of stress_level classes within both subsets.

While the three classes exhibited comparable frequencies and did not demonstrate any substantial imbalance, the pipeline integrated the RandomOverSampler as a regulated rebalancing approach within the training workflow. The operation was applied exclusively to the training data of each partition, following the relevant preprocessing steps and prior to the estimation tuning stage. In no case was oversampling applied to the external validation folds or the holdout set. This approach was implemented to avoid contamination of the evaluation and to preserve the independence of the performance estimates.

3.6. Algorithms and Hyperparameter Tuning

A comparative analysis was conducted on ensemble and boosting learning algorithms, encompassing bagging, boosting, voting, and meta-ensemble methods. The primary set of reported models encompassed AdaBoost, Gradient Boosting, Random Forest, Extra Trees, Bagging, Voting, Stacking, XGBoost, and LightGBM. The computational environment documented the availability of additional libraries, such as CatBoost. It is imperative that the inclusion of these libraries in the final tables corresponds exclusively to the models that were actually executed and documented.

The implementation of all models occurred within a unified pipeline that incorporated preprocessing, rebalancing, and the estimation process. This configuration guaranteed that the processes of imputation, scaling, one-hot encoding, and oversampling were applied exclusively to the training data within each partition.

Hyperparameter tuning was confined to the inner loop of the nested cross-validation, employing StratifiedKFold with three partitions. The evaluation metric employed was F1-weighted, a suitable choice for multiclass classification when performance is weighted by class frequency. GridSearchCV was utilized in smaller spaces, while RandomizedSearchCV with 50 iterations was employed in larger spaces. In both cases, random_state was set to 42 to ensure reproducibility. The hyperparameters that were explored included, depending on the algorithm, the number of estimators, the learning rate, the maximum depth, the number of leaves, subsampling, the proportion of variables, and classifier-specific parameters.

3.7. Nested Validation and Model Selection Criteria

The comparison between algorithms was performed using stratified nested cross-validation. The outer loop utilized 10 folds to estimate generalizable performance within the nested-train subset, while the inner loop employed 3 folds for hyperparameter tuning. In each outer fold, the entire

 doi:10.20944/preprints202606.0311.v1

pipeline was trained on the training partition, optimized internally, and evaluated on the external partition not used during tuning.

The performance of the model was quantified using multiclass metrics that focused on discrimination, class balance, and predictive agreement. The multiclass metrics included accuracy, balanced accuracy, F1-weighted, F1-macro, precision, recall, Cohen's Kappa, MCC, and multiclass ROC-AUC. ROC-AUC was calculated using a one-vs-rest scheme, with the weighted average primarily reported to maintain consistency with the primary selection metric. Due to the multiclass nature of the outcome, specifically binary metrics, such as binary precision-recall area under the curve (PR-AUC), binary expected calibration error (ECE), and decision curve analysis, were not used for the main interpretation.

The final selection was made in accordance with a predetermined hierarchical rule. The F1-weighted average was designated as the primary criterion, with MCC, ROC-AUC weighted, and cumulative computation time serving as tie-breaking criteria. This rule served to reduce the necessity for post hoc decisions and to elucidate the study's analytical priorities.

3.8. Statistical Comparison, Final Evaluation, and Calibration

To make a formal comparison of the relative performance of the algorithms, the nonparametric Friedman test was applied to the F1-weighted values obtained fold by fold in the outer loop of the nested cross-validation. Post hoc comparisons were conducted using Nemenyi and compact letter groupings exclusively when the overall test reached statistical significance; otherwise, the mean ranks were regarded as purely descriptive.

The algorithm that was selected according to the predefined hierarchical rule was retrained on the entire nested-train subset. The workflow for preprocessing, encoding, rebalancing, and hyperparameter tuning was maintained throughout the process. The internal holdout set remained isolated during the selection, optimization, and calibration processes, and was utilized solely for the final evaluation.

Probabilistic calibration was performed using CalibratedClassifierCV, with a sigmoid method and three-fold internal cross-validation, restricted to the nested-train subset. The RandomOverSampler remained encapsulated within the pipeline and was applied exclusively during the training of the base classifier on the training partitions, without affecting the validation folds or the holdout set. In each internal partition, the calibration function was estimated based on predictions generated on non-oversampled data. Finally, probabilistic quality was evaluated on the original holdout using the multiclass Brier score, calculated from the predicted probabilities and the one-hot encoding of the true class.

3.9. Considerations Regarding Rigor and Reproducibility

The pipeline was designed according to the principles of reproducibility, fair evaluation, and prevention of data leakage. The integration of imputation, standardization, coding, rebalancing, and estimation into a unified pipeline structure was implemented to prevent cross-contamination between the training and evaluation phases. Hyperparameter tuning was confined to the inner loop of the nested validation, while the internal holdout was maintained in isolation until the final stage.

The utilization of fixed random seeds, the documentation of the computational environment, and the recording of the versions of the main libraries employed were meticulously recorded. Furthermore, the training and evaluation time was documented as an ancillary performance metric. When considered as a whole, this architecture provides a robust, technically consistent methodological foundation that is aligned with contemporary publishing standards in applied ML.

3.10. Complementary Explainability Strategy

Given that post hoc explainability may be subject to technical limitations depending on the algorithm, the type of probabilistic output, and the multiclass structure of the problem, a detailed SHAP analysis was applied to the Gradient Boosting model, which was selected as the final classifier in accordance with the study's predefined hierarchical rule. The decision was predicated on three

 doi:10.20944/preprints202606.0311.v1

primary rationales. First, the Gradient Boosting algorithm demonstrated consistent performance metrics in the nested cross-validation framework, attaining the maximum F1-weighted score among the evaluated algorithms. Second, although the Friedman test did not reveal statistically significant differences between the models, Gradient Boosting maintained a competitive and methodologically consistent position within the group of highest-performing models. Third, the sequential tree-based architecture enables post hoc interpretation through the application of feature attribution techniques, thereby facilitating the analysis of the predictive patterns learned by the model.

4. Results

This section presents the empirical findings obtained using the ML pipeline designed for the multiclass prediction of student stress, with information leakage control. The analysis summarizes the structure and quality of the dataset, the comparative performance of the evaluated ensemble models, the statistical evidence supporting the model selection, the final evaluation on the holdout set, and the complementary interpretability results. The findings demonstrate that the proposed framework attained high and consistent predictive capabilities, though the top-performing algorithms exhibited analogous behavior, with no indication of discernible inferential superiority.

The analysis revealed no occurrence of duplicate records or missing values within the predictors or the outcome variable. The target variable, *stress_level*, exhibited a multiclass structure comprising three categories: 0, 1, and 2. The distribution of cases among these categories was nearly balanced, with 373, 358, and 369 cases, respectively. This distribution reduced the risk of bias associated with severe class imbalance and facilitated a more stable comparison among the evaluated algorithms. Following a methodological refinement of the *pipeline*, the final analytical set consisted of 3 numerical predictors and 17 explicitly coded categorical predictors, thereby confirming the consistency between the semantic nature of the variables and their statistical treatment.

The experimental partition yielded a *nested-train* subset of 880 observations and an independent *holdout* set of 220 observations, both of which were generated through stratification. This configuration enabled the dissociation of the model selection and optimization process from the final evaluation, thereby facilitating a more conservative and less biased estimate of generalizable performance.

4.1. Comparative Performance in Nested Cross-Validation

The application of nested cross-validation revealed that the top-performing models demonstrated high and relatively consistent overall performance. According to the study's primary metric, Gradient Boosting achieved the highest average performance, with an *accuracy* of 89.55 ± 3.16%, *balanced accuracy* of 89.52 ± 3.21%, F1-*weighted* of 89.54 ± 3.17%, MCC of 0.845 ± 0.047, and ROC-AUC *weighted* of 98.59 ± 0.92%. The findings suggest a high degree of discriminatory capability, adequate stability across different partitions, and a balanced classification pattern across the designated classes.

As shown in **Table 1, t**he XGBoost and LightGBM models demonstrated performance that was comparable to the optimal model. XGBoost attained an F1-*weighted* score of 89.43 ± 3.83%, an MCC of 0.842 ± 0.057, and a ROC-AUC *weighted* of 98.57 ± 0.99%, while LightGBM achieved an F1-*weighted* score of 89.27 ± 4.09%, an MCC of 0.841 ± 0.061, and a ROC-AUC *weighted* of 98.51 ± 1.05%. Indeed, the absolute differences between the three best-performing models were negligible, suggesting that the problem contains a robust predictive signal that can be consistently captured by different families of ensemble-based algorithms.

 doi:10.20944/preprints202606.0311.v1

Table 1. Comparative performance of algorithms in nested cross-validation (sorted by F1-weighted score).

Model	Accuracy (%)	Balanced accuracy (%)	F1-weighted (%)	MCC	ROC-AUC weighted (%)	Brier score	Total time (s)
Gradient Boosting	89.55 ± 3.16	89.52 ± 3.21	89.54 ± 3.17	0.845 ± 0.047	98.59 ± 0.92	0.1594 ± 0.0485	615.02
XGBoost	89.43 ± 3.83	89.41 ± 3.84	89.43 ± 3.83	0.842 ± 0.057	98.57 ± 0.99	0.1542 ± 0.0622	94.47
LightGBM	89.32 ± 4.09	89.29 ± 4.11	89.27 ± 4.09	0.841 ± 0.061	98.51 ± 1.05	0.1492 ± 0.0536	386.71
Voting	88.64 ± 4.67	88.61 ± 4.69	88.63 ± 4.69	0.831 ± 0.070	98.43 ± 1.09	0.1316 ± 0.0418	122.25
AdaBoost	88.41 ± 4.07	88.40 ± 4.10	88.43 ± 4.08	0.831 ± 0.061	98.05 ± 1.55	0.5433 ± 0.0189	257.11
Random Forest	88.41 ± 3.55	88.40 ± 3.58	88.40 ± 3.56	0.828 ± 0.052	98.37 ± 1.06	0.1248 ± 0.0357	161.25
Stacking	88.41 ± 4.44	88.37 ± 4.48	88.39 ± 4.46	0.828 ± 0.066	98.36 ± 1.13	0.1492 ± 0.0394	2451.79
Bagging	88.41 ± 3.74	88.37 ± 3.77	88.38 ± 3.73	0.827 ± 0.056	98.41 ± 1.06	0.1266 ± 0.0368	7.84
Extra Trees	87.39 ± 3.73	87.39 ± 3.73	87.41 ± 3.72	0.813 ± 0.056	98.43 ± 0.98	0.1240 ± 0.0355	135.51

From a computational efficiency perspective, significant discrepancies were identified between models with comparable predictive performance. In particular, XGBoost showed a marginal reduction compared to Gradient Boosting in F1-*weighted* (89.43 ± 3.83% versus 89.54 ± 3.17%), MCC (0.842 ± 0.057 versus 0.845 ± 0.047), and ROC-AUC *weighted* (98.57 ± 0.99% versus 98.59 ± 0.92%), but required considerably less total time (94.47 s versus 615.02 s). This behavior demonstrates a more favorable performance–computational cost ratio for XGBoost, especially in scenarios where time efficiency is an operational criterion of significance.

In a similar vein, Bagging exhibited competitive performance, with an F1-*weighted* score of 88.38 ± 3.73%, an MCC of 0.827 ± 0.056, and a ROC-AUC *weighted* of 98.41 ± 1.06%, utilizing a mere 7.84 s of total processing time. Despite its suboptimal performance in comparison to the leading models, its low computational cost renders it a potentially viable option for exploratory, iterative, or time-constrained applications. Taken together, these results suggest that, while Gradient Boosting was consistently selected in accordance with the predefined hierarchical rule, other algorithms exhibited favorable efficiency profiles that could be considered in contexts where reducing training time is a priority.

4.2. Formal Model Selection and Statistical Comparison

The model selection was executed in accordance with the a priori hierarchical criteria that had been specified beforehand. The mean F1-weighted score was utilized as the primary metric, with MCC, ROC-AUC weighted, and computational time serving as secondary tie-breaking criteria. According to the established protocol, Gradient Boosting was designated as the ultimate model, as it attained the highest mean F1-weighted score in the nested cross-validation procedure. However, it is important to note that this selection should not be interpreted as evidence of absolute statistical or practical superiority over the other algorithms. This is because the Friedman test did not identify significant differences among the evaluated models.

In particular, XGBoost and LightGBM exhibited a high degree of similarity to the selected model, with only minor discrepancies in F1-weighted, MCC, and ROC-AUC weighted scores. Moreover,

 doi:10.20944/preprints202606.0311.v1

XGBoost attained the highest average rank in the Friedman test, although Gradient Boosting achieved the highest overall mean F1-weighted score. This discrepancy is not unexpected, given that Friedman ranks are calculated on an individual basis, whereas the selection rule was based on predefined overall averages. Consequently, the findings indicate a scenario of statistical equivalence among the top-performing models, rather than the presence of a single algorithm that is distinctly dominant.

As shown in **Table 2**, Gradient Boosting should be understood as the model selected based on the predefined analytical criteria, not as an inferentially superior model. XGBoost and LightGBM are methodologically valid and statistically comparable alternatives, especially in contexts where computational efficiency, ranking stability, or operational feasibility carry greater weight in the final decision.

Table 2. Ranking of the final model selection based on the predefined hierarchical rule.

Rank	Model	F1-weighted average	MCC average	ROC-AUC weighted average	Total time (s)
1	Gradient Boosting	0.895403	0.845225	0.985948	615.02
2	XGBoost	0.894329	0.842304	0.985749	94.47
3	LightGBM	0.892704	0.841016	0.985059	386.71
4	Voting	0.886323	0.831309	0.984270	122.25
5	AdaBoost	0.884322	0.830861	0.980513	257.11

The nonparametric Friedman test, applied to the F1-weighted scores obtained from the external folds of the nested cross-validation, did not reveal any statistically significant differences between the evaluated algorithms ($\chi^2 = 10.953$; $p = 0.204$). As the overall Friedman test did not attain statistical significance, no confirmatory Nemenyi post hoc comparisons were conducted. Consequently, the reported mean ranks should be interpreted solely as an ordinal description of the relative performance observed in the external folds, and not as evidence of statistically significant differences between pairs of algorithms. This decision is predicated on the recognition that the global test did not demonstrate sufficient inferential separation, as evidenced by the presence of minor discrepancies among models that were not adequately addressed through conventional interpretation.

As shown in **Table 3**, this outcome aligns with the observed proximity among the top-performing models. In particular, Gradient Boosting, XGBoost, and LightGBM achieved very similar mean F1-weighted scores—0.895403, 0.894329, and 0.892704, respectively—as well as small differences in MCC and ROC-AUC weighted scores. Consequently, the selection of Gradient Boosting should be interpreted as a methodologically consistent decision with the a priori prioritization rule, rather than as evidence of a robust inferential advantage over the compared alternatives.

Table 3. Statistical comparison of algorithms: Friedman's average rank.

Model	F1-weighted average	SD	Average rank
XGBoost	0.894329	0.038301	3.35
LightGBM	0.892704	0.040937	4.05
Gradient Boosting	0.895403	0.031678	4.50
Voting	0.886323	0.046867	4.80
AdaBoost	0.884322	0.040831	5.00
Bagging	0.883807	0.037348	5.20

 doi:10.20944/preprints202606.0311.v1

Model	F1-weighted average	SD	Average rank
Random Forest	0.884032	0.035562	5.60
Stacking	0.883914	0.044634	5.65
Extra Trees	0.874056	0.037175	6.85

Note: Friedman's ranks are calculated fold-wise; a lower rank is better. They may differ from the order based on overall means. They are not inferential because the Friedman test was not significant.

4.3. Final Evaluation on the Holdout Set

The model that was ultimately selected, *GradientBoostingClassifier*, was retrained using the entire *nested-train* subset and the optimal hyperparameters that were identified during the tuning phase, which included *n_estimators* = 200 and *learning_rate* = 0.01. Subsequently, its predictive probabilities were calibrated using a sigmoid strategy, with the aim of improving the reliability of the probabilistic estimates prior to external evaluation.

In the independent *holdout* set, the model demonstrated an *accuracy* of 0.8818, a *balanced accuracy* of 0.8821, an F1-*weighted* score of 0.8818, an F1-*macro* score of 0.8819, an MCC of 0.8237, and a ROC-AUC weighted score of 0.9861. These results suggest that the model performs consistently and balances on data that were not utilized during the model selection or tuning process. Furthermore, the similarity between the performance observed in the *holdout* set and that obtained during nested cross-validation—mean F1-*weighted* of 0.8954 ± 0.0317 and mean MCC of 0.8452—suggests adequate generalization ability, with no evidence of a substantial degradation in performance outside the development phase.

The independent holdout evaluation reported in Table 4 confirms that the selected Gradient Boosting model maintained stable predictive performance outside the cross-validation procedure. Nevertheless, because the Friedman test did not detect statistically significant differences among the evaluated algorithms, these holdout results should be interpreted as evidence of external consistency for the model selected under the predefined hierarchical criteria, rather than as confirmation of inferential superiority over statistically comparable alternatives, particularly XGBoost and LightGBM.

Table 4. Performance of the final model on the independent holdout set.

Metrics	Value
Accuracy	0.881818
Balanced accuracy	0.882132
F1-weighted	0.881792
F1-macro	0.881938
Weighted accuracy	0.883877
Sensitivity/weighted recall	0.881818
Macro precision	0.883842
Sensitivity/recall (macro)	0.882132
ROC-AUC macro	0.986089
ROC-AUC weighted	0.986078
Cohen's Kappa	0.822746
MCC	0.823690
Brier score	0.137442

Note. The performance of the model was estimated using an independent holdout set. ROC-AUC values are reported as unweighted and weighted averages. The Brier score is a metric that summarizes the probabilistic calibration error, with lower values indicating better calibration.

 doi:10.20944/preprints202606.0311.v1

The final model exhibited robust and balanced predictive performance on the independent holdout set, with highly comparable values for accuracy, balanced accuracy, F1-weighted, and F1-macro. Additionally, the elevated values for ROC-AUC macro and weighted imply a high capacity to differentiate between classes, while Kappa and MCC signify substantial agreement that exceeds what would be expected by chance.

Table 5 presents the classification metrics by class, along with the unweighted and weighted averages. The support is equivalent to the number of observations that have been evaluated in each class within the independent holdout set.

Table 5. Performance by class of the final model on the independent holdout set.

Class	Precision	Sensitivity (recall)	F1-score	Support
0	0.93	0.84	0.88	74
1	0.88	0.92	0.90	72
2	0.85	0.89	0.87	74
Macro average	**0.88**	**0.88**	**0.88**	**220**
Weighted average	**0.88**	**0.88**	**0.88**	**220**

Note. Precision measures the proportion of correct predictions within each estimated class; sensitivity or recall indicates the proportion of actual cases in each class that are correctly identified; the F1-score summarizes the balance between accuracy and sensitivity. The macro average assigns equal weight to each class, while the weighted average takes into account the support of each class.

The classification pattern exhibited in the holdout set demonstrates balanced performance across the three classes, with no indications of substantial deterioration in any category. The F1-scores demonstrated stability within a narrow range, fluctuating between 0.87 and 0.90, thereby indicating the model's consistent predictive capability.

Class 1 demonstrated the strongest overall performance, attaining the highest F1-score (0.90) and exhibiting an optimal balance between precision and recall. Conversely, Class 2 exhibited the lowest F1-score (0.87), although this value remains within a range consistent with stable performance. Class 0 exhibited marginally diminished sensitivity relative to the other classes, yet there was an absence of evidence indicative of classification collapse or a significant diminution in discriminative power.

Additionally, the proximity between the overall and weighted averages suggests that the model's overall performance was not significantly influenced by any particular class. This outcome aligns with the almost equitable distribution of support observed in the holdout set, wherein the three classes exhibited comparable sizes. When considered as a whole, these findings lend support to the stability of the final model and suggest that its predictive performance was distributed in a consistent manner across the evaluated categories.

4.4. Probabilistic Quality and Interpretability

The sigmoid calibration applied to the final model yielded a *Brier score* of 0.1374 on the *holdout* set, suggesting adequate predictive performance for the multiclass context under analysis. This outcome aligns with the discriminative performance evident in the final evaluation, wherein the model attained substantial values for ROC-AUC *macro* and ROC-AUC *weighted*, nearing 0.986, accompanied by balanced performance across classes.

The analysis by class revealed a consistent classification pattern. Class 1 exhibited the highest F1 score of 0.90, followed by Class 0 with 0.88, and Class 2 with 0.87. Although class 0 achieved the

 doi:10.20944/preprints202606.0311.v1

highest precision value (0.93), its *recall* value was relatively lower (0.84), indicating a moderate tendency to classify this category more conservatively. In contrast, classes 1 and 2 achieved *recall* values of 0.92 and 0.89, respectively, suggesting adequate identification capability across all three categories. The equivalence between the *macro* and *weighted* averages—both equal to 0.88 for precision, *recall*, and F1-score—supports the absence of any significant bias associated with class support and confirms balanced performance on the independent set.

Prior to the presentation of the global and local explainability analyses, it is important to acknowledge the use of SHAP and LIME as complementary techniques for interpreting the predictive behavior of the final model. While SHAP enables the examination of variable contributions at both the global level and for individual predictions, LIME provides a local explanation centered on a specific observation. In this context, Figure 1 presents the SHAP beeswarm plot for the "high academic stress" class (stress_level = 2), allowing the identification of the variables with the greatest contribution to the model's predictions in a one-vs-rest multiclass classification scheme. However, these explanations should not be interpreted as causal evidence. Rather, they represent an interpretive decomposition of the associations learned by the Gradient Boosting model in the multiclass classification task of stress levels. This interpretation is particularly relevant because it enables the identification of variables that either increase or decrease the predictive evidence for specific classes. Additionally, it facilitates an assessment of whether the model's decisions are consistent with the psychological, physiological, social, and environmental patterns represented in the data.

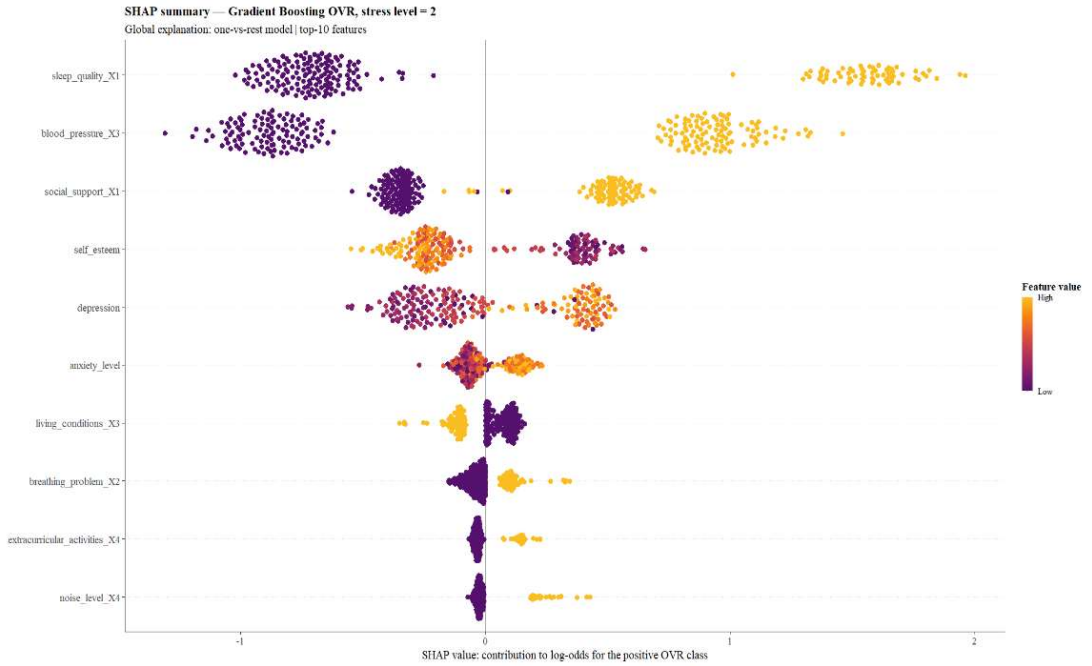

Figure 1. Beeswarm plot of SHAP values for the "high academic stress" class (stress_level = 2) in a one-vs-rest multiclass classification scheme using Gradient Boosting.

Note. The figure illustrates the 10 variables that exhibit the highest average local contribution to predicting the class stress level, which is designated as "stress_level = 2." Each data point signifies an individual observation, with its position on the horizontal axis denoting the effect of the SHAP value on the model's output on a log-odds scale. Positive values indicate an increase in the relative probability of belonging to the high-stress class, while negative values indicate a decrease. The color of the variable represents the magnitude of its original value, ranging from low to high. Overall, variables such as sleep_quality, blood_pressure, social_support, self_esteem, and depression show a significant influence on the classification of high stress levels. It is imperative to interpret these results as predictive associations learned by the model and not as causal relationships.

 doi:10.20944/preprints202606.0311.v1

The SHAP beeswarm plot demonstrates that the Gradient Boosting-OvR model exhibits its maximum discriminative power for the stress_level = 2 class in a restricted set of predictors associated with sleep quality, physiological indicators, social support, self-esteem, psychological symptoms, living conditions, and somatic manifestations. According to the ranking by absolute mean global impact, the variables with the greatest contribution to the prediction of this class were sleep_quality_X1, blood_pressure_X3, social_support_X1, self_esteem, depression, anxiety_level, living_conditions_X3, breathing_problem_X2, extracurricular_activities_X4, and noise_level_X4. This hierarchy indicates that the identification of elevated stress levels is not contingent upon a solitary predominant predictor. Rather, it is the result of a nonlinear amalgamation of psychological, physiological, social, and environmental factors that are imparted to the classifier.

The utilization of color gradients facilitates the interpretation of the direction of local contributions. In sleep_quality_X1, high values are predominantly concentrated in positive SHAP regions, suggesting that this category enhances the model's capacity to predict stress_level = 2. Conversely, low values are primarily situated in negative regions, thereby diminishing the support for that class. A similar pattern is observed in blood_pressure_X3, where elevated values tend to shift toward positive SHAP values, suggesting that this physiological category provides relevant predictive evidence for classifying individuals into the high stress level category.

In the case of social_support_X1, high values are primarily located in positive SHAP regions, while low values appear more frequently in negative or near-zero regions. This pattern suggests that the category encoded as X1 in social support contributes positively to the classification of the analyzed class. Conversely, self-esteem exhibits an inverse relationship, wherein low values are frequently linked to positive SHAP contributions, while high values are predominantly concentrated in negative regions. This finding indicates that diminished self-esteem levels are associated with heightened model support for the class, as evidenced by the stress_level = 2 outcome.

The psychological variables depression and anxiety_level demonstrate a greater degree of heterogeneity in their effects. The contributions of both variables are found to be significant, though with a considerable proportion of values tending toward zero. This observation indicates that their influence is contingent on the multivariate context in which they manifest, particularly in relation to other predictor variables. Conversely, elevated depression scores exhibited a discernible propensity for favorable contributions, aligning with their established predictive capability within high-stress profiles. Conversely, breathing_problem_X2 and noise_level_X4 demonstrate more localized contributions, albeit with positive shifts in select cases. This suggests that specific categories of somatic symptoms and environmental conditions may amplify the log-odds assigned to the positive class.

In a similar vein, living_conditions_X3 and extracurricular_activities_X4 exhibited effects of smaller overall magnitude, though these effects were not deemed to be insignificant. The contributions of these factors are found to be negligible, suggesting a more nuanced influence that may be contingent on interactions with more substantial variables, such as sleep quality, blood pressure, social support, self-esteem, and depression. This distribution is anticipated in boosting models, wherein predictive power may emerge from nonlinear combinations of variables as opposed to isolated marginal effects.

The local explainability analysis presented in Figure 2 indicates that the Gradient Boosting-OvR model identifies the class stress_level = 2 through a decision structure influenced by variables related to sleep, blood pressure, social support, self-esteem, psychological symptoms, living conditions, and somatic manifestations. However, because SHAP values represent a local decomposition of the model's output within a one-vs-rest binary framework, these observations must be interpreted exclusively as predictive associations and not as causal relationships. To enhance the interpretive robustness of the analysis, future work should complement this explanation with SHAP dependency curves, interaction analyses, stability assessment of importance scores under cross-validation, and verification of predictive performance on out-of-sample data.

 doi:10.20944/preprints202606.0311.v1

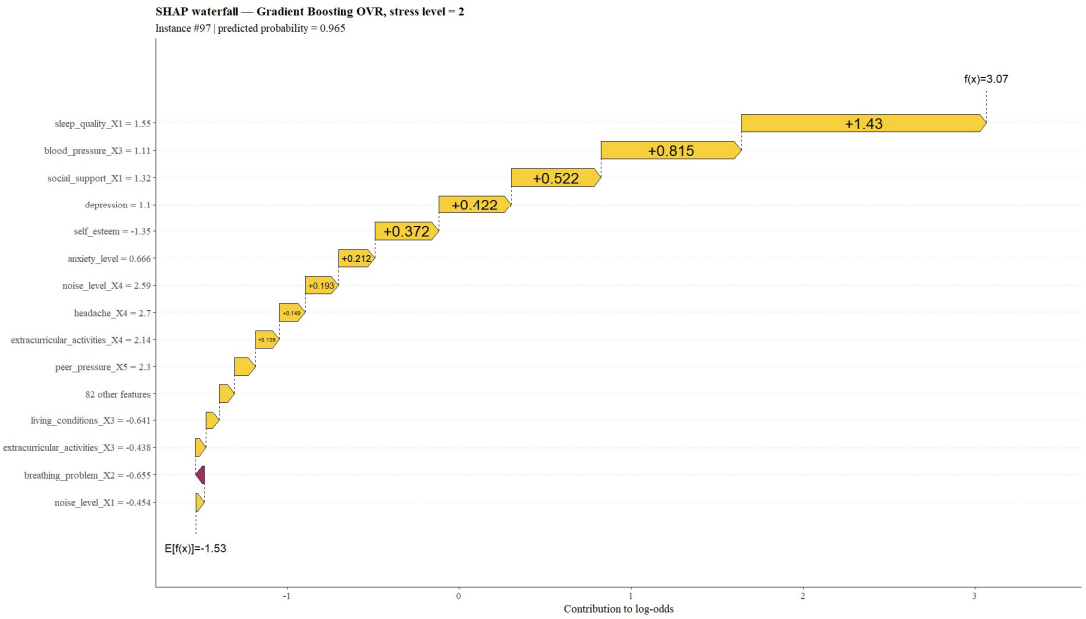

Figure 2. Waterfall plot of SHAP scores for a local prediction from the Gradient Boosting-OvR model for the stress_level = 2 class.

Note. The waterfall plot shows the additive decomposition of the individual prediction for instance #97, classified as positive (stress_level = 2), with a predicted probability of 0.965. The horizontal axis represents the contribution of each predictor to the output of the one-vs-rest binary classifier on a log-odds scale. The explanation starts from the model's overall expected value, E[f(x)] = −1.53, and successively accumulates the local contributions of the variables until reaching the final output f(x) = 3.07, reflecting strong predictive evidence in favor of the analyzed class.

The local explanation is dominated by a small set of high-magnitude positive contributions. The variable sleep_quality_X1 = 1.55 is the primary driver of the prediction, with an approximate contribution of +1.43 log-odds. Subsequently, the following variables were observed: blood_pressure_X3 = 1.11 (+0.815), social_support_X1 = 1.32 (+0.522), depression = 1.10 (+0.422), and self_esteem = −1.35 (+0.372). An additional positive contribution is also observed from anxiety_level = 0.666 (+0.212). Collectively, these predictors account for the majority of the observed shift from the model's baseline toward a prediction that strongly favors stress_level = 2.

In contrast, certain variables, including noise_level_X4 = 2.59, headache_X4 = 2.7, and extracurricular_activities_X4 = 2.14, exert a moderating effect of lesser magnitude. These variables contribute negatively to the evidence supporting the positive classification, thereby reducing its overall strength. However, these effects do not counterbalance the cumulative impact of the predominant predictors, resulting in a definitive outcome that remains distinctly positioned toward the high-stress category.

From a methodological perspective, this representation should be interpreted as a local, additive, and non-causal explanation of the model's prediction. SHAP values quantify the impact of the observed values in this instance on the classifier's predictive evidence regarding the class designated as "stress_level = 2" in relation to the other categories. However, it is imperative to note that SHAP values do not imply direct causal relationships. Consequently, the interpretation of this figure must be supplemented with global analyses—such as beeswarm diagrams, SHAP dependency curves, interaction analysis, and stability assessment via cross-validation—to support a more robust characterization of the model's behavior.

Figure 3 presents a LIME-based local explanation for an instance at the 85th confidence quantile, evaluated using a Gradient Boosting model calibrated for the multiclass prediction of stress level. Panel (a) illustrates the calibrated probabilities assigned to each category. The analyzed instance was

 doi:10.20944/preprints202606.0311.v1

classified as stress level = 1, with a calibrated probability of 99.2%, while the probabilities corresponding to classes 0 and 2 were practically zero. This pattern suggests a prediction that is predominantly concentrated within the intermediate stress class and distant from the decision boundary in the probability space of the calibrated model.

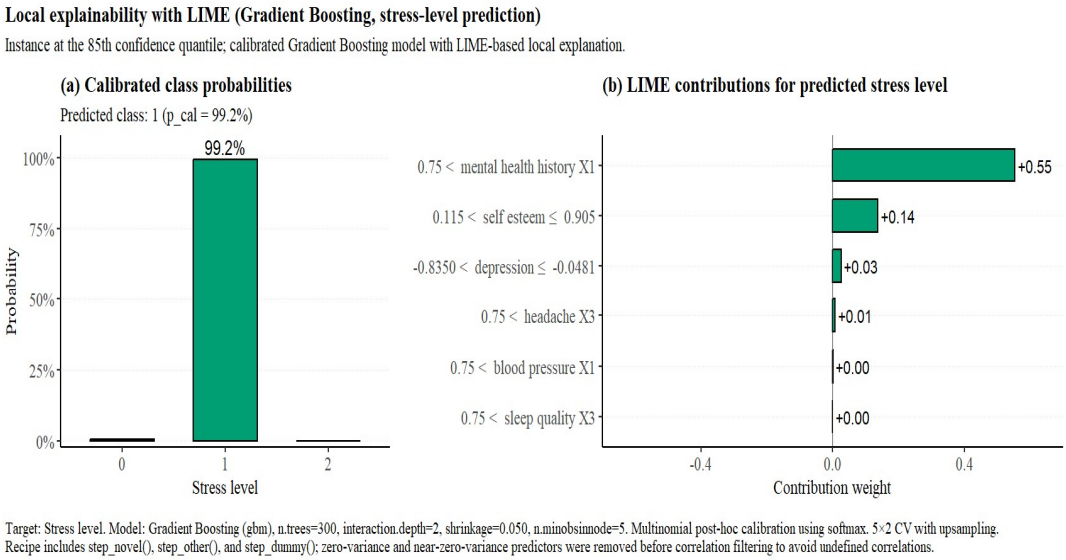

Figure 3. Local explainability using LIME for stress level prediction with calibrated gradient boosting.

Panel (b) presents the local contributions estimated by LIME for the predicted class. According to the local linear approximation, the primary factor contributing to the classification as stress level = 1 is mental_health_history_X1 > 0.75, with an approximate weight of +0.55. This indicates that this condition provides the majority of the local evidence in favor of the predicted class. Second, the interval 0.115 < self_esteem ≤ 0.905 contributes positively with an approximate weight of +0.14. To a more moderate extent, the interval −0.8350 < depression ≤ −0.0481 makes a positive contribution of approximately +0.03.

The variables headache_X3 > 0.75, blood_pressure_X1 > 0.75, and sleep_quality_X3 > 0.75 also appear in the local explanation, although with weights that are nearly negligible. This finding indicates that, for this particular observation, these predictors were selected by LIME as part of the explanatory neighborhood. However, their marginal influence on the final decision was negligible compared to the influence of mental_health_history_X1 and self_esteem. Consequently, the local classification is dominated by a small number of conditions, rather than by a homogeneous contribution from all the predictors shown.

From a methodological perspective, LIME weights should not be interpreted as overall coefficients of the Gradient Boosting model. Rather, they should be interpreted as parameters of a local linear approximation constructed around the instance being analyzed. Consequently, the interpretation should be constrained to this particular observation. In a similar manner, the thresholds indicated within the labels, such as 0.75 < mental_health_history_X1 or 0.115 < self_esteem ≤ 0.905, are derived from the discretization applied by LIME to variables that underwent transformation, coding, or standardization during the preprocessing stage. Consequently, these findings should not be interpreted as clinical, diagnostic, or psychometric cutoff points.

In summary, the prediction for this instance (stress level = 1) is primarily supported by the local contribution of mental_health_history_X1, reinforced to a lesser extent by self_esteem and depression. The high probability estimate indicates a consistent decision for this observation; however, the LIME explanation remains local, approximate, and non-causal. Consequently, these results should be interpreted as instance-specific predictive evidence and supplemented with global significance analyses, SHAP explanations, stability assessment under resampling, and validation on out-of-sample data.

 doi:10.20944/preprints202606.0311.v1

5. Discussion

The findings of this study demonstrate that ensemble models constitute a robust strategy for the multiclass prediction of student stress levels based on psychometric, academic, somatic, and contextual variables. In the context of nested cross-validation, the performance of the evaluated algorithms was found to be consistently high, with accuracy values ranging from 87.39% to 89.55%, F1-weighted values between 87.41% and 89.54%, MCC ranging from 0.813 to 0.845, and ROC-AUC weighted values exceeding 98%. Gradient Boosting demonstrated the optimal mean performance, closely followed by XGBoost and LightGBM. This finding indicates that the predictive structure of the problem was consistently captured by diverse families of tree-based ensembles.

However, Friedman's nonparametric test did not reveal any statistically significant differences among the models evaluated, $\chi^2 = 10.953$; $p = 0.204$. Therefore, although Gradient Boosting was selected as the final model according to the predefined hierarchical criteria—F1-weighted, MCC, ROC-AUC weighted, and computational time—its superiority should be interpreted as a descriptive and methodologically consistent advantage, but not as conclusive inferential evidence of algorithmic dominance. This finding is significant because XGBoost achieved nearly equivalent performance with lower computational cost. This suggests that, in real-world educational settings, model selection should consider not only accuracy but also efficiency, calibration, interpretability, and feasibility of implementation.

The evaluation of the proposed pipeline's stability was confirmed through an independent holdout set. The calibrated model demonstrated an accuracy of 0.8818, a balanced accuracy of 0.8821, an F1-weighted score of 0.8818, an MCC of 0.8237, and a ROC-AUC weighted of 0.9861. The proximity of these results to those obtained in the nested cross-validation suggests that there has been no substantial degradation outside the development phase. Additionally, the similarity between the macro and weighted metrics indicates balanced performance across categories, with no indications of classification collapse in a particular category. The Brier score of 0.1374 indicates sufficient probabilistic quality for a multiclass task; however, this calibration should be reevaluated prior to transferring the model to new cohorts or institutions.

These results are partially at odds with the systematic review by Daza et al. [42], in which SVM and logistic regression emerge as frequently used and effective techniques for predicting anxiety and stress among college students. In this study, however, tree-based ensembles and meta-ensembles achieved levels of performance that are highly competitive. This discrepancy can be attributed to the heterogeneous nature of the analyzed variables. Ensemble models exhibit greater flexibility in capturing the nonlinear relationships between psychological, academic, physiological, and social factors.

The performance obtained in this study is comparable to that reported by Tariq et al. [43], who observed accuracies of 0.89 for Random Forest, 0.87 for XGBoost, and 0.85 for Gradient Boosting in the classification of university stress levels. These accuracies were obtained using variables similar to those included in this study, such as blood pressure, perceived safety, sleep quality, teacher–student relationship, and extracurricular participation. In a similar vein, Rois et al. [55] reported an accuracy close to 0.90 and an AUC of approximately 0.87 with Random Forest, identifying heart rate, blood pressure, sleep, and lifestyle habits as relevant predictors. The findings from this study are consistent with those obtained elsewhere, thereby underscoring the importance of somatic and behavioral indicators in the modeling of student stress.

Although certain studies have documented superior performance—for instance, Anand et al. [33], with an accuracy of 93.48% and an F1-score of 93.14% in academic stress classification, or Pereira et al. [45], with accuracies approaching 95–97% in burnout risk profiles—such comparisons must be interpreted with caution. A significant proportion of extant studies address binary tasks or simpler classification structures; however, this study focuses on a multiclass problem with heterogeneous variables and potentially greater discriminative complexity. Concurrently, studies such as those by Campanella et al. [49], Chen and Lee [56], and Almadhor et al. [47] employ high-resolution physiological signals, including PPG, ECG, EEG, or EDA, in conjunction with deep or optimized architectures. This methodological framework limits the direct comparability of the findings with an approach based on self-reports, academic variables, and contextual factors.

 doi:10.20944/preprints202606.0311.v1

In this regard, an accuracy of approximately 88–90%, accompanied by an MCC greater than 0.82 and a ROC-AUC weighted close to 0.986, represents solid performance for a multiclass task involving student stress. Moreover, these results exceed those documented in studies that employ questionnaires or psychosocial variables in complex mental health problems. For instance, Hasan et al. [18] reported accuracies ranging from 0.57 to 0.69 for the prediction of depression, anxiety, and stress in contexts characterized by high prevalence and increased clinical complexity. The findings indicate that the incorporation of psychometric, academic, somatic, and contextual variables enables a significant enhancement in discriminative power, obviating the necessity for physiological sensors or complex architectures.

A methodological contribution of this study is the incorporation of a rigorous comparative evaluation using nested cross-validation, hyperparameter optimization, probabilistic calibration, holdout evaluation, and formal statistical testing between algorithms. As demonstrated in the literature, reviews by Mittal et al. [15] and Mahajan et al. [51] underscore the efficacy of strategies such as stacking and voting across various clinical domains. However, a significant limitation of these comparisons is the utilization of aggregated metrics without the implementation of robust inferential tests. In contrast, the findings of this study demonstrate that, under controlled conditions, the disparities between complex models and classical ensembles can be mitigated. This finding indicates that the perceived superiority of specific meta-ensembles may be contingent on the validation protocol, preprocessing techniques, and the configuration of the dataset.

Moreover, the findings of this study are consistent with those reported by Martinović et al. [30], Villar and De Andrade [57], and Tang et al. [58], who emphasize the competitive performance of boosting methods and optimized ensembles on metrics such as accuracy, F1-score, and ROC-AUC. However, the findings of this study also demonstrate the necessity of incorporating computational cost as a critical criterion in decision-making processes. It is noteworthy that models with highly similar performance metrics can exhibit significant disparities in terms of training time. This has substantial practical ramifications for the implementation of screening or early-warning systems within educational institutions that are constrained by limited computational resources.

Explanatory analyses serve to reinforce the substantive plausibility of the model. The SHAP analysis revealed that the prediction of stress levels was predominantly influenced by variables such as sleep quality, self-esteem, anxiety, depressive symptoms, basic needs, extracurricular activities, blood pressure, headaches, concerns about future career, and teacher–student relationships. This pattern aligns with the findings of de Filippis and Al Foysal [59], who identified self-esteem, sleep quality, and anxiety as pivotal predictors of stress in students, encompassing psychological, physiological, environmental, academic, and social dimensions. This finding aligns with the observations made by Abdul Rahman et al. [60], who underscored the significance of lifestyle variables and academic performance as predictors of negative mental well-being. Additionally, it is consistent with the findings of Pereira et al. [45], who highlighted the pivotal role of psychological distress, emotional regulation, and health habits in the development of burnout.

A salient methodological consideration pertains to the fact that the detailed explainability was not derived directly from the final calibrated Gradient Boosting model, but rather from a comparable XGBoost model. This decision does not invalidate the interpretive analysis; rather, it imposes limitations on its scope. Given that XGBoost demonstrated nearly equivalent performance and that the statistical comparison revealed no significant differences between models, its explanations can be considered a reasonable approximation of the dominant predictive patterns of the problem. However, it is imperative to interpret the explanatory results as complementary and exploratory in nature, as they do not substitute for direct explanatory validation of the final model nor permit the establishment of causal relationships.

Local explanations, as elucidated by SHAP waterfall and LIME, serve to complement the overarching interpretation by demonstrating how specific combinations of variables result in a prediction that is shifted toward a particular class. Specifically, local explanations demonstrated that factors such as sleep quality, blood pressure, headaches, history of mental health issues, social support, self-esteem, and depression can substantially influence the model's output in individual cases. These results align with studies incorporating explainability techniques in mental health and

 doi:10.20944/preprints202606.0311.v1

education, such as Xia et al. [61], who used SHAP to interpret LightGBM models in contexts of stress and reading performance, as well as Shanto and Jony [52], Villar and De Andrade [57], Pujadas et al. [44], Geng et al. [53], and Tariq et al. [43], who applied SHAP and, in some cases, LIME to increase the transparency of predictive models in stress, mental health, academic dropout, and adaptation to online learning environments.

From an applied perspective, the consistent identification of variables related to sleep, self-esteem, anxiety, depressive symptoms, blood pressure, underlying health conditions, extracurricular activities, and teacher–student relationships suggest priority areas for psychoeducational interventions. These findings can serve as a guide for programs focused on sleep hygiene, building self-efficacy, academic support, personalized tutoring, and improving the institutional environment. From an applied perspective, these results should not be interpreted as sufficient evidence to implement an early detection system or longitudinal monitoring. The cross-sectional design of the study prevents the assessment of temporal changes, future incidence, or preventive capacity. Instead, the model should be regarded as a methodological instrument for the classification of observed levels of stress and the exploration of associated predictive profiles. Its eventual use in institutional screening systems would require prospective validation, performance evaluation in new cohorts, and ethical-operational analysis regarding the use of predictive probabilities in real educational contexts.

It is imperative to acknowledge the limitations of the study when interpreting the findings. Although the holdout set provides an independent evaluation within the same dataset, external validation is required across new institutions, cohorts, and sociocultural contexts to determine the model's generalizability. Additionally, the probabilistic calibration of the system may be influenced by distributional shifts, particularly in cases where the system is implemented in populations exhibiting disparate stress profiles. Future research should integrate multimodal sources of psychometric, academic, behavioral, and physiological information. Rigorous validation protocols, formal statistical comparison between models, and multilayer explainability frameworks should be maintained. This approach will facilitate progress toward more accurate, interpretable, generalizable, and ethically acceptable decision support systems for the management of student stress and mental health.

6. Conclusions

The findings of this study demonstrate that the multiclass classification of observed levels of student stress can be addressed with high performance using tree-based ensemble models. Gradient Boosting achieved the best average performance according to the predefined hierarchical criterion; however, XGBoost and LightGBM achieved very similar results, and the Friedman test did not reveal any statistically significant differences among the evaluated algorithms. Consequently, the selection of Gradient Boosting should be interpreted as a methodological decision consistent with the prioritization protocol. However, this does not imply conclusive inferential superiority over the other models.

From a methodological perspective, the study provides a multi-criteria evaluation framework that combines nested cross-validation, holdout evaluation, statistical comparison, computational cost analysis, global probabilistic calibration, and complementary explainability. This methodological framework enables the evaluation of models not solely on their predictive performance but also on their stability, efficiency, probabilistic reliability, and interpretability. Specifically, SHAP and LIME analyses conducted on a comparable XGBoost model indicate that variables such as sleep quality, self-esteem, anxiety, depressive symptoms, blood pressure, basic needs, extracurricular activities, and teacher–student relationships offer pertinent predictive insights. However, it is imperative to acknowledge that these attributions should be regarded as complementary predictive explanations rather than as causal or diagnostic evidence, nor as a direct explanation of the final Gradient Boosting model.

It is imperative to interpret the findings with caution due to the utilization of public, observational, and cross-sectional data, devoid of external validation or longitudinal follow-up. Consequently, the model should not yet be regarded as a proven system for early detection,

 doi:10.20944/preprints202606.0311.v1

monitoring, or prevention. Rather, it should be considered a methodological approach for classifying observed levels of stress in a specific dataset. In the future, researchers should validate the pipeline in external cohorts and different institutions. They should also more accurately document the psychometric validity of the instruments. Researchers should evaluate multiclass calibration by class. They should analyze sensitivity to coding and rebalancing decisions. Researchers should incorporate multimodal sources. Then, they should consider institutional implementation in real-world psychoeducational support settings.

Author Contributions: D.A.G.: writing—review and editing, writing—original draft, visualization, resources, methodology, investigation, formal analysis, conceptualization. W.M.R.: writing—original draft, methodology, investigation, formal analysis, project administration, supervision. M.Z.R.: writing—review and editing, methodology, investigation, supervision, resources. A.N.A.: conceptualization, funding acquisition, resources, visualization, validation. E.S.R.: funding acquisition, visualization, validation, supervision. M.A.L.V.: project administration, funding acquisition, resources, validation.

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset analyzed in this study is publicly available and can be accessed from the original source cited in the manuscript. These data were used for preprocessing, training, validation, and comparative evaluation of the machine learning models. No new primary data were generated as part of this research.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Douwes, R.; Metselaar, J.; Pijnenborg, G.H.M.; Boonstra, N. Well-Being of Students in Higher Education: The Importance of a Student Perspective. *Cogent Educ* **2023**, *10*, 2190697, doi:10.1080/2331186X.2023.2190697.
2. Arif, N.M.N.A.; Roslan, N.S.; Ismail, S.B.; Nayak, R.D.; Jamian, M.R.; Mohamad Ali Roshidi, A.S.; Edward, T.C.; Kamal, M.A.; Mohd Amin, M.M.; Shaari, S.; et al. Prevalence and Associated Factors of Psychological Distress and Burnout among Medical Students: Findings from Two Campuses. *Int J Environ Res Public Health* **2021**, *18*, 8446, doi:10.3390/ijerph18168446.
3. March-Amengual, J.-M.; Cambra Badii, I.; Casas-Baroy, J.-C.; Altarriba, C.; Comella Company, A.; Pujol-Farriols, R.; Baños, J.-E.; Galbany-Estragués, P.; Comella Cayuela, A. Psychological Distress, Burnout, and Academic Performance in First Year College Students. *Int J Environ Res Public Health* **2022**, *19*, 3356, doi:10.3390/ijerph19063356.
4. Pascoe, M.C.; Hetrick, S.E.; Parker, A.G. The Impact of Stress on Students in Secondary School and Higher Education. *Int J Adolesc Youth* **2020**, *25*, 104–112, doi:10.1080/02673843.2019.1596823.
5. Smith, M.D.; Wesselbaum, D. Global Evidence on the Prevalence of and Risk Factors Associated with Stress. *J Affect Disord* **2025**, *374*, 179–183, doi:10.1016/j.jad.2025.01.053.
6. Li, H. Multicultural Data Assistance Mining Analysis for Ideological and Political Education in Smart Education Platforms Using Artificial Intelligence. *Wireless Netw* **2025**, *31*, 567–581, doi:10.1007/s11276-024-03772-8.
7. Gustavson, K.; Knudsen, A.K.; Nesvåg, R.; Knudsen, G.P.; Vollset, S.E.; Reichborn-Kjennerud, T. Prevalence and Stability of Mental Disorders among Young Adults: Findings from a Longitudinal Study. *BMC Psychiatry* **2018**, *18*, 65, doi:10.1186/s12888-018-1647-5.
8. Tan, W.; Chen, L.; Zhang, Y.; Xi, J.; Hao, Y.; Jia, F.; Hall, B.J.; Gu, J.; Wang, S.; Lin, H.; et al. Regional Years of Life Lost, Years Lived with Disability, and Disability-Adjusted Life-Years for Severe Mental Disorders in Guangdong Province, China: A Real-World Longitudinal Study. *Glob Health Res Policy* **2022**, *7*, 17, doi:10.1186/s41256-022-00253-3.

 doi:10.20944/preprints202606.0311.v1

9. Ullah, H.; Arbab, S.; Liu, C.; Du, Q.; Khan, S.A.; Khan, S.; Dad, O.; Tian, Y.; Li, K. Source of Stress-Associated Factors Among Medical and Nursing Students: A Cross-Sectional Study. *J Nurs Manag* **2025**, *2025*, 9928649, doi:10.1155/jonm/9928649.
10. Conley, C.S.; Durlak, J.A.; Dickson, D.A. An Evaluative Review of Outcome Research on Universal Mental Health Promotion and Prevention Programs for Higher Education Students. *J Am Coll Health* **2013**, *61*, 286–301, doi:10.1080/07448481.2013.802237.
11. Ahorsu, D.K.; Sánchez Vidaña, D.I.; Lipardo, D.; Shah, P.B.; Cruz González, P.; Shende, S.; Gurung, S.; Venkatesan, H.; Duongthipthewa, A.; Ansari, T.Q.; et al. Effect of a Peer-led Intervention Combining Mental Health Promotion with Coping-strategy-based Workshops on Mental Health Awareness, Help-seeking Behavior, and Wellbeing among University Students in Hong Kong. *Int J Ment Health Syst* **2021**, *15*, 6, doi:10.1186/s13033-020-00432-0.
12. Roshan, R.; Hamid, S.; Kumar, R.; Hamdani, U.; Naqvi, S.; Zill-e-Huma; Adeel, U. Utilizing the CFIR Framework for Mapping the Facilitators and Barriers of Implementing Teachers Led School Mental Health Programs – A Scoping Review. *Soc Psychiatry Psychiatr Epidemiol* **2025**, *60*, 535–548, doi:10.1007/s00127-024-02762-7.
13. Zhang, Z.; Chen, H.; Ye, Y.; Chen, H.; Guo, H.; Zhou, J. Entropy-Based Risk Network Identification in Adolescent Self-Injurious Behavior Using Machine Learning and Network Analysis. *Transl Psychiatry* **2025**, *15*, 299, doi:10.1038/s41398-025-03511-3.
14. Zhang, Z. Early Warning Model of Adolescent Mental Health Based on Big Data and Machine Learning. *Soft Comput* **2024**, *28*, 811–828, doi:10.1007/s00500-023-09422-z.
15. Mittal, S.; Mahendra, S.; Sanap, V.; Churi, P. How Can Machine Learning Be Used in Stress Management: A Systematic Literature Review of Applications in Workplaces and Education. *Int J Inf Manag Data Insights* **2022**, *2*, 100110, doi:10.1016/j.jjimei.2022.100110.
16. Geronimo, S.M.; Hernandez, A.A.; Abisado, M.B. Academic Stress of Students in Higher Education Using Machine Learning: A Systematic Literature Review. In Proceedings of the 2023 IEEE 13th International Conference on System Engineering and Technology (ICSET); October 2023; pp. 141–146.
17. Arias, E.M.; Parraga-Alava, J.; Montenegro, D.Z. Stress Detection among Higher Education Students: A Comprehensive Systematic Review of Machine Learning Approaches. In Proceedings of the 2024 Tenth International Conference on eDemocracy & eGovernment (ICEDEG); June 2024; pp. 1–8.
18. Hasan, M.E.; Arif, M.; Rakibul Hasan, S.M.; Muwanguzi, M.; Abaatyo, J.; Kaggwa, M.M.; ALmerab, M.M.; Atroszko, P.A.; Muhit, M.; Al-Mamun, F.; et al. Prevalence, Associated Factors, and Machine Learning-Based Prediction of Depression, Anxiety, and Stress among University Students: A Cross-Sectional Study from Bangladesh. *J Health Popul Nutr* **2025**, *44*, 361, doi:10.1186/s41043-025-01095-8.
19. Wang, Y.; Lu, S.; Harter, D. Towards Collaborative and Intelligent Learning Environments Based on Eye Tracking Data and Learning Analytics: A Survey. *IEEE Access* **2021**, *9*, 137991–138002, doi:10.1109/ACCESS.2021.3117780.
20. Barreto, A.; Zhai, J.; Adjouadi, M. Non-Intrusive Physiological Monitoring for Automated Stress Detection in Human-Computer Interaction. In Proceedings of the Human–Computer Interaction; Lew, M., Sebe, N., Huang, T.S., Bakker, E.M., Eds.; Springer: Berlin, Heidelberg, 2007; pp. 29–38.
21. Broglia, E.; Barkham, M. Adopting the Principles and Practices of Learning Health Systems in Universities and Colleges: Recommendations for Delivering Actionable Data to Improve Student Mental Health. *Cogent Mental Health* **2024**, *3*, 2301339, doi:10.1080/28324765.2023.2301339.
22. Umematsu, T.; Sano, A.; Taylor, S.; Picard, R.W. Improving Students' Daily Life Stress Forecasting Using LSTM Neural Networks. In Proceedings of the 2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI); May 2019; pp. 1–4.
23. Ding, Y.; Chen, X.; Fu, Q.; Zhong, S. A Depression Recognition Method for College Students Using Deep Integrated Support Vector Algorithm. *IEEE Access* **2020**, *8*, 75616–75629, doi:10.1109/ACCESS.2020.2987523.
24. Skrbinjek, V.; Dermol, V. Predicting Students' Satisfaction Using a Decision Tree. *Tert Educ Manag* **2019**, *25*, 101–113, doi:10.1007/s11233-018-09018-5.

 doi:10.20944/preprints202606.0311.v1

25. Leevy, J.L.; Khoshgoftaar, T.M.; Bauder, R.A.; Seliya, N. A Survey on Addressing High-Class Imbalance in Big Data. *J Big Data* **2018**, *5*, 42, doi:10.1186/s40537-018-0151-6.
26. Hasanin, T.; Khoshgoftaar, T.M.; Leevy, J.L.; Bauder, R.A. Severely Imbalanced Big Data Challenges: Investigating Data Sampling Approaches. *J Big Data* **2019**, *6*, 107, doi:10.1186/s40537-019-0274-4.
27. Li, Z.; Kamnitsas, K.; Glocker, B. Overfitting of Neural Nets Under Class Imbalance: Analysis and Improvements for Segmentation. In Proceedings of the Medical Image Computing and Computer Assisted Intervention – MICCAI 2019; Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P.-T., Khan, A., Eds.; Springer International Publishing: Cham, 2019; pp. 402–410.
28. Carrington, A.M.; Manuel, D.G.; Fieguth, P.W.; Ramsay, T.; Osmani, V.; Wernly, B.; Bennett, C.; Hawken, S.; Magwood, O.; Sheikh, Y.; et al. Deep ROC Analysis and AUC as Balanced Average Accuracy, for Improved Classifier Selection, Audit and Explanation. *IEEE Trans Pattern Anal Mach Intell* **2023**, *45*, 329–341, doi:10.1109/TPAMI.2022.3145392.
29. Chicco, D.; Warrens, M.J.; Jurman, G. The Matthews Correlation Coefficient (MCC) Is More Informative Than Cohen's Kappa and Brier Score in Binary Classification Assessment. *IEEE Access* **2021**, *9*, 78368–78381, doi:10.1109/ACCESS.2021.3084050.
30. Martinović, M.; Dokic, K.; Pudić, D. Comparative Analysis of Machine Learning Models for Predicting Innovation Outcomes: An Applied AI Approach. *Appl Sci* **2025**, *15*, 3636, doi:10.3390/app15073636.
31. Abu-Shaira, M.; Shi, W. Unveiling Statistical Significance of Online Regression Over Multiple Datasets. In Proceedings of the 2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR); August 2024; pp. 274–279.
32. Carvalho, D.V.; Pereira, E.M.; Cardoso, J.S. Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics* **2019**, *8*, 832, doi:10.3390/electronics8080832.
33. Anand, R.V.; Md, A.Q.; Urooj, S.; Mohan, S.; Alawad, M.A.; C., A. Enhancing Diagnostic Decision-Making: Ensemble Learning Techniques for Reliable Stress Level Classification. *Diagnostics* **2023**, *13*, 3455, doi:10.3390/diagnostics13223455.
34. Saklani, S.; Manchanda, M.; Bisht, G.; Thapa, K. Comparing Random Forest and XGBoost for Sentiment Classification of Student Social Media Posts: A Case Study on Pre-Board Exam Stress. In Proceedings of the 2025 Global Conference in Emerging Technology (GINOTECH); May 2025; pp. 1–6.
35. Nnadi, L.C.; Watanobe, Y.; Rahman, M.M.; John-Otumu, A.M. Prediction of Students' Adaptability Using Explainable AI in Educational Machine Learning Models. *Appl Sci* **2024**, *14*, 5141, doi:10.3390/app14125141.
36. Islam, M.T.; Ashraf, K.; Hosen, Md.H.; Nawar, S.; Asgar, S. Predictive Modeling of Anxiety Levels in Bangladeshi University Students: A Voting-Based Approach with LIME and SHAP Explanations. In Proceedings of the 2024 International Conference on Advances in Computing, Communication, Electrical, and Smart Systems (iCACCESS); March 2024; pp. 01–06.
37. Hooshyar, D.; Yang, Y. Problems With SHAP and LIME in Interpretable AI for Education: A Comparative Study of Post-Hoc Explanations and Neural-Symbolic Rule Extraction. *IEEE Access* **2024**, *12*, 137472–137490, doi:10.1109/ACCESS.2024.3463948.
38. A. Marouf, A.; F. Ashrafi, A.; Ahmed, T.; Emon, T. A Machine Learning Based Approach for Mapping Personality Traits and Perceived Stress Scale of Undergraduate Students. *IJMECS* **2019**, *11*, 42–47, doi:10.5815/ijmecs.2019.08.05.
39. Flesia, L.; Monaro, M.; Mazza, C.; Fietta, V.; Colicino, E.; Segatto, B.; Roma, P. Predicting Perceived Stress Related to the Covid-19 Outbreak through Stable Psychological Traits and Machine Learning Models. *JCM* **2020**, *9*, 3350, doi:10.3390/jcm9103350.
40. Naghavi, A.; Teismann, T.; Asgari, Z.; Mohebbian, M.R.; Mansourian, M.; Mañanas, M.Á. Accurate Diagnosis of Suicide Ideation/Behavior Using Robust Ensemble Machine Learning: A University Student Population in the Middle East and North Africa (MENA) Region. *Diagnostics* **2020**, *10*, 956, doi:10.3390/diagnostics10110956.
41. Liu, Y.; Xie, Y.-N.; Li, W.-G.; He, X.; He, H.-G.; Chen, L.-B.; Shen, Q. A Machine Learning-Based Risk Prediction Model for Post-Traumatic Stress Disorder during the COVID-19 Pandemic. *Medicina* **2022**, *58*, 1704, doi:10.3390/medicina58121704.

 doi:10.20944/preprints202606.0311.v1

42. Daza, A.; Saboya, N.; Necochea-Chamorro, J.I.; Zavaleta Ramos, K.; Vásquez Valencia, Y.D.R. Systematic Review of Machine Learning Techniques to Predict Anxiety and Stress in College Students. *Inform Med Unlocked* **2023**, *43*, 101391, doi:10.1016/j.imu.2023.101391.
43. Tariq, R.; Orozco-del-Castillo, M.G.; Zamir, M.T.; Ramírez-Montoya, M.S.; Wilberforce, T. Explainable Artificial Intelligence for Predictive Modeling of Student Stress in Higher Education. *Sci Rep* **2025**, *15*, 38375, doi:10.1038/s41598-025-22171-3.
44. Pujadas, E.R.; Díaz-Caneja, C.M.; Stevanovic, D.; Quintero, M.F.; Martín-Isla, C.; Hernández-González, J.; Atehortúa, A.; Lazrak, N.; Pries, L.; Delespaul, P.; et al. Longitudinal Prediction of Mental Health Outcomes in Vulnerable Youth Using Machine Learning. *Cogn Comput* **2025**, *17*, 152, doi:10.1007/s12559-025-10509-y.
45. Pereira, M.G.; Santos, M.; Magalhães, R.; Rodrigues, C.; Araújo, O.; Durães, D. Burnout Risk Profiles in Psychology Students: An Exploratory Study with Machine Learning. *Behav Sci* **2025**, *15*, 505, doi:10.3390/bs15040505.
46. Younis, E.M.G.; Zaki, S.M.; Kanjo, E.; Houssein, E.H. Evaluating Ensemble Learning Methods for Multi-Modal Emotion Recognition Using Sensor Data Fusion. *Sensors* **2022**, *22*, 5611, doi:10.3390/s22155611.
47. Almadhor, A.; Sampedro, G.A.; Abisado, M.; Abbas, S. Efficient Feature-Selection-Based Stacking Model for Stress Detection Based on Chest Electrodermal Activity. *Sensors* **2023**, *23*, 6664, doi:10.3390/s23156664.
48. Hadhri, S.; Hadiji, M.; Labidi, W. A Voting Ensemble Classifier for Stress Detection. *J Inf Telecommun* **2024**, *8*, 399–416, doi:10.1080/24751839.2024.2306786.
49. Campanella, S.; Altaleb, A.; Belli, A.; Pierleoni, P.; Palma, L. A Method for Stress Detection Using Empatica E4 Bracelet and Machine-Learning Techniques. *Sensors* **2023**, *23*, 3565, doi:10.3390/s23073565.
50. Haghish, E.F.; Nes, R.B.; Obaidi, M.; Qin, P.; Stänicke, L.I.; Bekkhus, M.; Laeng, B.; Czajkowski, N. Unveiling Adolescent Suicidality: Holistic Analysis of Protective and Risk Factors Using Multiple Machine Learning Algorithms. *J Youth Adolescence* **2024**, *53*, 507–525, doi:10.1007/s10964-023-01892-6.
51. Mahajan, P.; Uddin, S.; Hajati, F.; Moni, M.A. Ensemble Learning for Disease Prediction: A Review. *Healthcare* **2023**, *11*, 1808, doi:10.3390/healthcare11121808.
52. Shanto, S.S.; Jony, A.I. Interpretable Ensemble Learning Approach for Predicting Student Adaptability in Online Education Environments. *Knowledge* **2025**, *5*, 10, doi:10.3390/knowledge5020010.
53. Geng, S.; Wang, J.; Xia, Y.; Niu, B.; Deng, X.; Wu, X. Predicting Generalized Anxiety Disorder among Chinese Depressed Adolescents: An Explainable Machine Learning Approach. *BMC Med Inform Decis Mak* **2025**, *25*, 406, doi:10.1186/s12911-025-03225-y.
54. Liu, C.; Yu, S. Students' Stress Prediction and Explainable Analysis Based on Improved Decision Trees. *Front Psychol* **2026**, *16*, doi:10.3389/fpsyg.2025.1684529.
55. Rois, R.; Ray, M.; Rahman, A.; Roy, S.K. Prevalence and Predicting Factors of Perceived Stress among Bangladeshi University Students Using Machine Learning Algorithms. *J Health Popul Nutr* **2021**, *40*, 50, doi:10.1186/s41043-021-00276-5.
56. Chen, Q.; Lee, B.G. Deep Learning Models for Stress Analysis in University Students: A Sudoku-Based Study. *Sensors* **2023**, *23*, 6099, doi:10.3390/s23136099.
57. Villar, A.; De Andrade, C.R.V. Supervised Machine Learning Algorithms for Predicting Student Dropout and Academic Success: A Comparative Study. *Discov Artif Intell* **2024**, *4*, 2, doi:10.1007/s44163-023-00079-z.
58. Tang, B.; Li, S.; Zhao, C. Predicting the Performance of Students Using Deep Ensemble Learning. *J Intell* **2024**, *12*, 124, doi:10.3390/jintelligence12120124.
59. De Filippis, R.; Foysal, A.A. Comprehensive Analysis of Stress Factors Affecting Students: A Machine Learning Approach. *Discov Artif Intell* **2024**, *4*, 62, doi:10.1007/s44163-024-00169-6.
60. Abdul Rahman, H.; Kwicklis, M.; Ottom, M.; Amornsriwatanakul, A.; H. Abdul-Mumin, K.; Rosenberg, M.; Dinov, I.D. Machine Learning-Based Prediction of Mental Well-Being Using Health Behavior Data from University Students. *Bioengineering* **2023**, *10*, 575, doi:10.3390/bioengineering10050575.
61. Xia, Z.; Lee, C.E.; Chen, C.-H.; Kuo, J.-Y.; Lim, K.Y.H. Mental States and Cognitive Performance Monitoring for User-Centered e-Learning System: A Case Study. In *Advances in Transdisciplinary Engineering*; Moser, B.R., Koomsap, P., Stjepandić, J., Eds.; IOS Press, 2022 ISBN 978-1-64368-338-6.

 doi:10.20944/preprints202606.0311.v1

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.s