



Explainable XGBoost Early-Warning Framework for Academic Stress-Based Student Mental Health Risk Mapping

Supriyono¹, Heru Noviyanto Firmansyah², Soni Adiyono³

¹²³Department of Information Systems, Faculty of Engineering, Universitas Muria Kudus, Kudus, Indonesia

*supriyono.si.umk@gmail.com

Abstract

Existing university mental health monitoring often depends on voluntary help-seeking or manual questionnaire interpretation, which may delay early support for students experiencing academic stress. This study proposes an explainable XGBoost-based early-warning framework for non-clinical mapping of student mental health risk from academic stress indicators. The single-site dataset comprised 1,002 anonymized student records from Universitas Muria Kudus. K-Means clustering was used to transform DASS-21 depression, anxiety, and stress scores into low, moderate-, and high-risk categories, while XGBoost predicted the cluster-derived labels using seven single-item academic stress indicators and engineered aggregate and interaction features. On a stratified hold-out testing set of 201 records, the model achieved weighted precision, recall, and F1-score values of 0.8907, 0.8905, and 0.8906, respectively, with class-level F1-scores of 0.9109 for low risk, 0.8900 for moderate risk, and 0.8713 for high risk. Additional ablation, clustering sensitivity, subgroup, threshold, and SHAP stability analyses were conducted to strengthen robustness and interpretability. The findings show that cumulative academic stress and interaction features involving parental expectations, exam anxiety, and learning-method adaptation were consistently influential predictors. The framework is intended to support early institutional prioritization and counseling referral, not clinical diagnosis. Generalization remains limited by the single-institution sample and the use of single-item academic stress indicators; therefore, local retraining and recalibration are required before institutional deployment, including implementation of the Streamlit prototype.

Keywords: academic stress; DASS-21; explainable AI; K-Means clustering; mental health risk; XGBoost

How to Cite: Supriyono, H. N. Firmansyah, and S. Adiyono, "Explainable XGBoost Early-Warning Framework for Academic Stress-Based Student Mental Health Risk Mapping", *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 10, no. 4, pp. 866 - 877, Jul. 2026.

Permalink/DOI: <https://doi.org/10.29207/resti.v10i4.7812>

Received: May 8, 2026

Accepted: June 21, 2026

Available Online: July 19, 2026

This is an open-access article under the [CC BY 4.0 License](#)

Published by [Ikatan Ahli Informatika Indonesia](#)

1. Introduction

Student mental health has become a strategic issue in higher education because psychological distress can reduce academic engagement, academic achievement, persistence, and social functioning [1]. University students face cumulative pressure from lectures, assignments, examinations, peer comparison, time-management demands, and expectations from themselves and others [2], [3]. These stressors may intensify symptoms of depression, anxiety, and stress when institutional monitoring is delayed or relies only on voluntary help-seeking [1], [2].

The Depression, Anxiety, and Stress Scale-21 (DASS-21) is widely used to measure depression, anxiety, and stress as three related but distinct negative emotional

states [3-5]. In educational settings, DASS-21 can support structured early screening, yet manual interpretation becomes less efficient when institutions must process hundreds or thousands of student records [6]. A purely rule-based interpretation may also fail to capture latent risk profiles and interaction patterns among academic stressors [7], [8].

Previous studies have applied machine learning and predictive modeling to mental health and stress-related screening problems, including fuzzy logic-based prediction, DASS-21-based classification, perceived stress modeling, and explainable mental health prediction [6], [9-12]. Clustering has been used to identify psychological profiles without predefined labels [8], [13], whereas supervised learning models

such as Random Forest, Support Vector Machine, Logistic Regression, and XGBoost have been used to classify risk outcomes [14-16]. However, many studies either classify mental health severity using predefined categories or generate clusters without converting those empirical groups into a reusable prediction model for new student records. This creates a methodological gap between latent risk discovery and operational early-warning prediction.

Compared with prior studies that relied on predefined DASS-21 severity categories or reported machine-learning performance without explanation stability, the present study positions its contribution as an applied, interpretable, and robustness-tested workflow. It does not claim algorithmic novelty in K-Means or XGBoost; rather, its novelty lies in separating DASS-21-based label construction from academic-stress-based prediction, testing engineered stress features through ablation, and evaluating interpretability through SHAP rank and bootstrap stability.

This study addresses that gap by separating label construction and prediction. DASS-21 subscale scores are used only to construct empirical mental health risk categories through K-Means clustering, while academic stress indicators are used as non-clinical predictors in the supervised XGBoost model. This separation reduces the risk of merely reproducing DASS-21 scoring rules and allows the model to estimate risk from academic stress patterns when full DASS-21 responses are not yet available [17].

XGBoost was selected because boosted tree models are particularly suitable for structured tabular data and can capture nonlinear relationships and interaction effects without assuming linearity among predictors [18], [19]. This is relevant because academic stress dimensions rarely operate in isolation. For example, parental expectations may intensify exam anxiety, and difficulty adapting to teaching methods may increase the effect of perceived academic failure. XGBoost was therefore compared with Random Forest, SVM, Decision Tree, and Logistic Regression to determine whether its nonlinear modeling capacity produced practical advantages.

Explainability is essential in mental health-related decision support because predictions may influence student support, counseling referral, and institutional trust [20], [21]. Black-box outputs are difficult to justify in educational environments, particularly when false positives may burden counseling units and false negatives may delay assistance for students requiring follow-up [22], [23]. Therefore, this study integrates SHAP explanations to clarify which academic stress features contribute most strongly to predicted risk, while explicitly limiting system use to non-clinical screening and prioritization.

The academic stress variables were coded computationally as acd1 to acd7. These codes do not represent a separate patented ACD instrument. They are

operational feature names for seven single-item academic stress indicators conceptually aligned with academic stress literature [3], [24]: learning-method adaptation stress (acd1), parental expectation-related stress (acd2), exam anxiety (acd3), perceived risk of academic failure (acd4), academic competition pressure (acd5), assignment load pressure (acd6), and difficulty understanding course material (acd7). Because single-item indicators have measurement limitations, this study treats them as operational screening signals rather than full psychometric constructs.

The contribution of this study is fourfold. First, it proposes a DASS-21-based hybrid K-Means and XGBoost framework for mapping mental health risk from academic stress indicators [25]. Second, it evaluates aggregate and interaction features through ablation analysis. Third, it expands validation using clustering sensitivity, repeated-initialization stability, subgroup performance, threshold analysis, and calibration diagnostics. Fourth, it uses SHAP summary, rank stability, and bootstrap-based feature importance to strengthen interpretability claims. The objective is to develop and evaluate a non-clinical early-warning framework that can support university-level prioritization of academic mentoring and counseling referral

2. Methods

2.1. Research Design and Data Context

This study used a quantitative data mining design based on the Cross-Industry Standard Process for Data Mining (CRISP-DM), consisting of business understanding, data understanding, preparation, modeling, evaluation, and deployment planning [26], [27]. The business understanding stage defined the institutional need for early mental health risk mapping. The data understanding stage examined DASS-21 scores and academic stress indicators. The preparation stage cleaned the dataset, normalized variables, constructed cluster labels, and generated engineered features. The modeling stage applied K-Means clustering and supervised classification, while the evaluation stage assessed predictive performance, interpretability, and robustness.

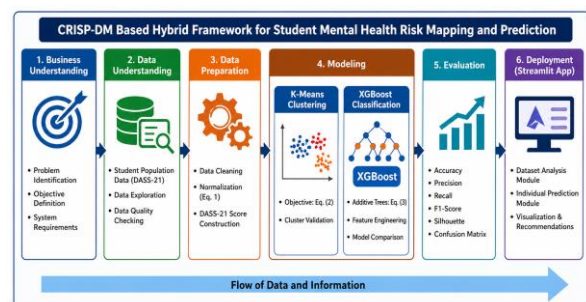


Figure 1. CRISP-DM-based hybrid framework for student mental health risk mapping and prediction. DASS-21 scores are used only for label construction, while academic stress indicators and engineered features are used for supervised prediction.

Figure 1 is provided to clarify the end-to-end analytical workflow before the individual data preparation, modeling, and evaluation procedures are described. The workflow clarifies the separation between risk-label construction and supervised prediction. This separation is central to the study because DASS-21 scores define the risk groups, while academic stress features are used to predict those groups without direct target leakage.

2.2. Dataset, Participants, and Ethical Boundaries

The dataset consisted of 1,002 anonymized student records from active students at Universitas Muria Kudus (UMK), covering the Faculty of Engineering and the Faculty of Economics and Business. The sample included students from Informatics Engineering, Information Systems, and Management programs, with data collected from December 2024 to January 2025. Students were included when they were actively enrolled, completed the DASS-21 and academic stress

measurement blocks, and agreed to the use of anonymized responses for research purposes.

Records with duplicate respondent identifiers, impossible score ranges, or incomplete key variables were removed before analysis. Personal identifiers were replaced with anonymous respondent codes before modeling. Administrative permission to collect the data was obtained from the relevant faculties at UMK. Because the study uses mental health-related data, the model output must be interpreted only as non-clinical decision support, not as diagnostic judgment [21-23]. The predictions should support academic mentoring, counseling referral, or further professional assessment, but should not be used to label students permanently, impose academic sanctions, or restrict academic opportunities. Table 1 provides a simplified view of the anonymized dataset structure used for modeling and evaluation.

Table 1. Sample structure of the student mental health dataset

No.	Respondent ID	Gender	Study Program	Cohort	DEP1	DEP2	...	ANX1	...	STR7	DEP Score	ANX Score	STR Score	MHR Score
1	R001	Female	Information Systems	2022	1	0	...	1	...	1	8	7	9	24
2	R002	Male	Informatics Engineering	2022	3	2	...	2	...	3	15	14	16	45
3	R003	Male	Informatics Engineering	2022	1	1	...	0	...	1	7	6	8	21
4	R004	Male	Informatics Engineering	2022	1	2	...	1	...	2	10	9	11	30
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1002	R1002	Female	Management	2023	1	0	...	0	...	0	5	4	6	15

2.3. Measures and Feature Construction

The DASS-21 block contained depression, anxiety, and stress items. Each item was scored on a four-point scale from 0 to 3, following the standard response format for the previous week. The depression, anxiety, and stress subscale scores were calculated by summing the seven items in each subscale and multiplying the result by 2 to maintain comparability with the full DASS framework [5]. These subscale scores were used only to construct risk labels and were excluded from the supervised predictor set to avoid target leakage [17].

The academic stress block contained seven single-item indicators on a five-point Likert scale. These indicators were used as practical early-warning signals rather than full latent constructs. Engineered features included

acd_mean to represent cumulative academic stress intensity, acd_std to represent variability across academic stressors, acd_max to represent the strongest stressor, and pairwise interaction terms such as acd2_x_acd3 to capture combined stress effects. The pairwise interactions were retained because the influence of academic pressure may arise from combinations of stressors rather than isolated dimensions [2], [24]. To reduce ambiguity in interpretation, the single-item academic stress indicators were not treated as independent psychometric subscales. Instead, they were positioned as operational screening variables that capture students' perceived academic pressures in specific learning situations.

Table 2. Academic stress dimensions used in the model.

Dimension (code)	Number of items	Scale	Example indicator
Learning-method adaptation stress (acd1)	1	5-point Likert	Difficulty adapting to the lecturer's teaching method
Parental expectation-related stress (acd2)	1	5-point Likert	Worry that academic grades may not match parental expectations
Exam anxiety (acd3)	1	5-point Likert	Restlessness or difficulty sleeping because of examinations
Perceived risk of academic failure (acd4)	1	5-point Likert	Anxiety about the possibility of repeating a course after failing an exam
Academic competition pressure (acd5)	1	5-point Likert	Perceived intense competition for academic achievement among peers
Assignment load pressure (acd6)	1	5-point Likert	Stress caused by too many assignments with short deadlines
Difficulty in understanding course material (acd7)	1	5-point Likert	Confusion or difficulty understanding the course material delivered in class

This distinction is important because the model does not attempt to validate a new academic stress instrument, but to examine whether practical academic stress signals can predict empirically derived mental health risk groups. Therefore, feature-importance interpretation focuses on risk-monitoring relevance rather than construct-level psychological diagnosis. Table 2 defines the academic stress indicators used as supervised predictors so that the computational feature codes can be interpreted substantively.

2.4. Preprocessing, Label Construction, and Reproducibility Protocol

Data cleaning included duplicate checking, missing-value inspection, validation of score ranges, and complete-case filtering for key DASS-21 and academic stress variables. DASS-21 item scores were validated within the 0-3 range, while acd1-acd7 were validated within the 1-5 Likert range. Missing or invalid responses were excluded when they could distort either DASS-21 score construction or academic stress feature engineering.

Because K-Means is distance-based, Min-Max normalization was applied to the DASS-21 subscale matrix before clustering. The supervised predictor matrix was constructed separately from acd1-acd7 and engineered academic stress features. DEP, ANX, STR, and the descriptive MHR score were not used as predictors in XGBoost. This separation was used to avoid target leakage and to ensure that the supervised model learned risk-related patterns from academic stress indicators rather than variables involved in label construction [17].

The Min-Max transformation used before clustering is expressed in Equation 1. This step ensured that the depression, anxiety, and stress subscales contributed comparably to Euclidean-distance-based clustering.

$$x' = (x - \min(x)) / (\max(x) - \min(x)) \quad (1)$$

K-Means clustering was applied to the normalized DASS-21 depression, anxiety, and stress subscale profile. The baseline random seed was set to 42. Candidate k values from 2 to 10 were evaluated using WCSS, silhouette score, Davies-Bouldin index, and repeated-initialization stability. Although internal compactness favored k = 2, the final k = 3 taxonomy was retained because it provides low-, moderate-, and high-risk groups that are more actionable for non-clinical triage and early support prioritization [22], [28].

The K-Means objective minimized within-cluster variation, as shown in Equation 2, where C_k denotes cluster k, x_i denotes a student vector, and μ_k denotes the centroid of cluster k.

$$J = \sum_{k=1}^K \sum_{x_i \in C_k} \|x_i - \mu_k\|^2 \quad (2)$$

After label construction, a stratified 80:20 train-test split was applied to preserve class proportions, producing 801 training records and 201 testing records. The supervised models were evaluated under the same

target labels, predictor set, and data partition. Five-fold stratified cross-validation was used to obtain comparative mean $\pm$ standard deviation metrics, while the hold-out test set was used for detailed class-wise performance, confusion matrix, ROC, PR, threshold, and error analyses. No oversampling procedure such as SMOTE was applied in the final model; instead, stratification and class-wise metrics were used to monitor imbalance effects [29], [30].

The main evaluation metrics were accuracy, precision, recall, and F1-score. Because false negatives in high-risk screening are practically important, recall and high-risk false-negative rate were emphasized alongside weighted F1-score.

2.5. Model Training, Ablation, Subgroup Analysis, and SHAP Stability

XGBoost was configured for multiclass classification and compared with Random Forest, Support Vector Machine, Decision Tree, and Logistic Regression. The comparative evaluation used stratified five-fold cross-validation on the same cluster-derived target labels and academic stress predictor set. The XGBoost configuration was kept consistent across the final pipeline, and model comparison was interpreted as an empirical benchmark rather than a claim of algorithmic novelty.

To answer the reviewer's reproducibility concern, the model comparison was conducted as a fixed-configuration benchmark rather than an exhaustive hyperparameter-optimization study. No automated Grid Search, Random Search, or Bayesian optimization was applied in the final reported experiment. All models were evaluated using the same stratified five-fold cross-validation folds, the same train-test partition, and random seed 42. Because the cluster-derived classes were sufficiently represented in the split, the final XGBoost model did not use SMOTE or class-specific sample weighting; imbalance effects were monitored through class-wise recall, high-risk false-negative rate, and subgroup performance.

The final static XGBoost configuration used `n_estimators = 100`, `learning_rate = 0.10`, `max_depth = 3`, `objective = multi:softprob`, `num_class = 3`, `eval_metric = mlogloss`, `subsample = 1.00`, `colsample_bytree = 1.00`, `min_child_weight = 1`, `gamma = 0`, `reg_alpha = 0`, `reg_lambda = 1`, `random_state = 42`, and `n_jobs = -1`. No automated hyperparameter search was performed in the reported experiment; all non-reported arguments followed the implementation defaults used during analysis. These values are reported to guarantee complete reproducibility of the fixed-configuration benchmark.

The XGBoost prediction function is summarized in Equation 3. The model combines multiple decision trees to form an additive ensemble, allowing nonlinear and interaction-based academic stress patterns to contribute to the final class prediction.

$$y_i = \sum_{t=1}^T f_t(x_i), f_t \in F \quad (3)$$

Ablation analysis was performed to quantify the relative contribution of feature engineering and model selection. The evaluated scenarios included baseline acd1-acd7 predictors, the addition of aggregate features, the addition of interaction features, the proposed XGBoost configuration, Random Forest baseline, and Logistic Regression baseline. This analysis directly assesses whether engineered stress features and nonlinear tree-based models provide improvements over simpler specifications.

Subgroup analysis was conducted to examine whether risk class distributions and supervised model performance varied across gender, study program, faculty, and cohort. Subgroups with adequate test-set support were evaluated using accuracy, F1-score, high-risk recall, and high-risk false-negative rate. These results were not interpreted as definitive fairness certification, but as an internal bias-screening step to identify groups where missed high-risk cases may require further monitoring.

SHAP analysis was used to explain feature contributions for the high-risk class and the multiclass model output [20], [31]. To avoid overinterpreting a single split, SHAP robustness was assessed using rank stability across five cross-validation folds and bootstrap-based feature importance stability. SHAP values were interpreted as model attribution rather than causal evidence; causal claims would require longitudinal or experimental research designs.

3. Results and Discussions

3.1 Cluster Construction, Sensitivity, and Stability

The three-cluster solution produced ordered increases in depression, anxiety, and stress means across the low-, moderate-, and high-risk groups. In raw DASS-21 score terms, C1 showed the lowest profile (DEP = 13.35, ANX = 11.08, STR = 7.34), C2 represented a moderate burden profile (DEP = 26.53, ANX = 23.50, STR = 18.59), and C3 represented the highest burden profile (DEP = 39.83, ANX = 35.55, STR = 30.43). This ordered pattern supports the substantive interpretation of the clusters as non-clinical risk categories rather than arbitrary numerical partitions. Table 3 reports the centroid profile of the selected three-cluster solution in raw DASS-21 score units.

Table 3. Cluster profile of the selected three-cluster solution in raw DASS-21 score units.

Cluster	DEP mean	ANX mean	STR mean	Interpretation
C1	13.35	11.08	7.34	Low Risk
C2	26.53	23.50	18.59	Moderate Risk
C3	39.83	35.55	30.43	High Risk

The profile in Table 3 shows a clear raw-score progression from C1 to C3, so the cluster labels are substantively ordered. This supports using the clusters as low-, moderate-, and high-risk categories for non-clinical prioritization.

Reporting the centroid values in raw DASS-21 score units also makes the cluster interpretation easier to understand for non-technical readers and institutional stakeholders. This is important because the model is intended to support practical risk prioritization, not only statistical grouping. Table 4 presents the clustering metrics used to compare candidate k values before selecting the operational three-cluster taxonomy.

Table 4. Clustering evaluation metrics for candidate k values using normalized DASS-21 subscale profiles.

k	WCSS	Silhouette Score
2	186.83	0.3819
3	129.49	0.3332
4	107.65	0.2748
5	98.24	0.2266
6	91.49	0.2178
7	86.78	0.2055
8	82.06	0.1961
9	77.56	0.2005
10	72.64	0.1967

Table 4 confirms that $k = 2$ produced the strongest silhouette value, but $k = 3$ still retained acceptable separation while offering a more useful three-level triage taxonomy. The selected solution therefore reflects a practical trade-off between statistical compactness and institutional interpretability.

The sensitivity analysis confirms the trade-off noted by the reviewer. The silhouette score was highest at $k = 2$, indicating stronger internal compactness, whereas $k = 3$ provided a more operationally useful taxonomy for differentiating transitional or moderate-risk students from high-risk students. The Davies-Bouldin trend also shows that cluster validity declines for larger k values, suggesting that additional groups would add fragmentation without improving interpretability. Therefore, the choice of $k = 3$ was based on a balance between internal validity and institutional actionability rather than on silhouette score alone. Figure 2 visualizes the sensitivity of cluster validity across candidate k values to justify the operational selection of three risk groups.

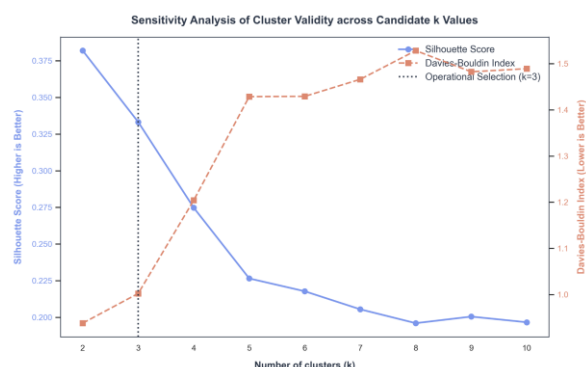


Figure 2. Cluster validity sensitivity across candidate k values.

Figure 2 shows that selecting $k = 3$ does not simply follow the highest silhouette value. Instead, the figure supports the argument that the final clustering decision balances statistical evidence with the practical need for three actionable risk levels [32-34].

K-Means stability was also evaluated across repeated random initializations. K-Means has also been used in previous student mental health clustering studies, supporting its relevance for exploratory risk profiling in educational datasets [35]. The stability plot indicates that silhouette values remained close to the baseline value, while Adjusted Rand Index (ARI) values were generally high, suggesting that the three-cluster solution was not driven by a single random initialization. This supports the robustness of the label construction step while still acknowledging that cluster-derived labels remain sample-dependent and require external validation. Figure 3 is included to show whether the selected K-Means solution remains stable when repeated with different random initializations.

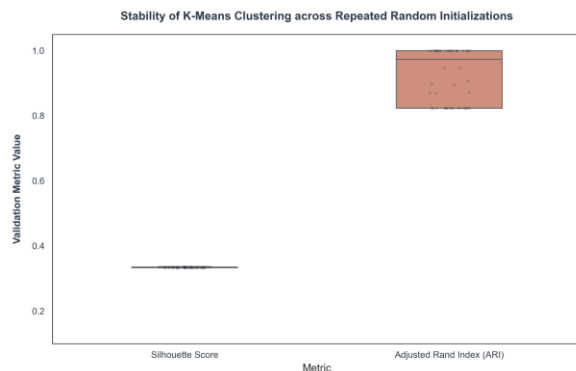


Figure 3. K-Means stability across repeated random initializations.

The stability result indicates that the selected three-cluster labels were not produced only by a favorable seed. The silhouette values remained stable, and ARI values were generally high, supporting the reproducibility of the clustering stage within the current dataset. Figure 4 complements the numerical cluster profiles by projecting the DASS-21 risk groups into a two-dimensional PCA space.

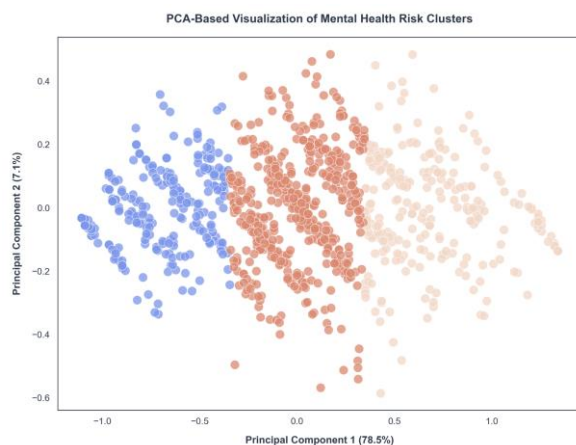


Figure 4. PCA visualization of the three mental health risk clusters.

The PCA projection shows ordered but partially overlapping risk regions, which is expected because mental health risk is continuous rather than perfectly discrete. The overlap also explains why most classification errors occur between adjacent risk classes [36].

3.2 Predictive Performance, Ablation, and Error Analysis

The proposed XGBoost model achieved balanced performance on the 201-record hold-out testing set. The low-risk class produced precision, recall, and F1-score values of 0.9184, 0.8824, and 0.9000. The moderate-risk class produced 0.8889, 0.8889, and 0.8889, while the high-risk class produced 0.8679, 0.9020, and 0.8846. The weighted-average precision, recall, and F1-score were 0.8910, 0.8905, and 0.8906, respectively. These values indicate that the model performed consistently across the three risk categories, with the high-risk class retaining strong recall on the hold-out test set. Table 5 summarizes the class-wise hold-out performance of the proposed XGBoost model.

Table 5. Class-wise hold-out performance of the XGBoost model on the 201-record testing set.

Class	Support	Precision	Recall	F1-score
Low Risk	51	0.9184	0.8824	0.9000
Moderate Risk	99	0.8889	0.8889	0.8889
High Risk	51	0.8679	0.9020	0.8846
Weighted average	201	0.8910	0.8905	0.8906

Table 5 indicates that the model performed consistently across classes, with the highest F1-score for low risk and a slightly lower F1-score for high risk. This pattern shows good overall performance but also identifies high-risk recall as the main point for operational monitoring. Figure 5 summarizes the hold-out classification errors so that both correct predictions and adjacent-class misclassifications can be evaluated.

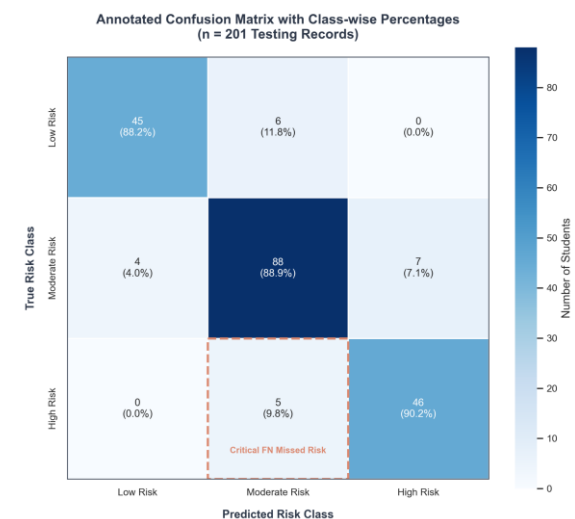


Figure 5. Annotated confusion matrix on the 201-record testing set.

which represent the most important missed-risk pattern for early-warning use.

The annotated confusion matrix shows that the model correctly classified 179 of 201 testing records. It correctly identified 45 low-risk students, 88 moderate-risk students, and 46 high-risk students. Misclassifications were concentrated between adjacent risk categories: six low-risk cases were predicted as moderate risk, four moderate-risk cases were predicted as low risk, seven moderate-risk cases were predicted

as high risk, and five high-risk cases were predicted as moderate risk. No low-risk case was predicted directly as high risk, and no high-risk case was predicted directly as low risk. This pattern is important because the most severe cross-extreme errors were absent, although high-risk false negatives remain practically important.

The comparison indicates that XGBoost and Random Forest achieved very similar overall performance, while SVM, Decision Tree, and Logistic Regression were lower. This supports the use of tree-based nonlinear models for academic stress predictors. Table 6 reports cross-validation-based model comparison results to benchmark XGBoost against alternative classifiers.

Table 6. Model performance comparison using stratified cross-validation mean $\pm$ standard deviation.

Model	Accuracy	Precision	Recall	F1-score
XGBoost	0.8603 $\pm$	0.8631 $\pm$	0.8603 $\pm$	0.8601 $\pm$
(Proposed)	0.0180	0.0176	0.0180	0.0180
Random	0.8583 $\pm$	0.8656 $\pm$	0.8583 $\pm$	0.8579 $\pm$
Forest	0.0103	0.0104	0.0103	0.0103
SVM	0.8423 $\pm$	0.8608 $\pm$	0.8423 $\pm$	0.8405 $\pm$
	0.0162	0.0163	0.0162	0.0164
Decision	0.8303 $\pm$	0.8325 $\pm$	0.8303 $\pm$	0.8304 $\pm$
Tree	0.0119	0.0121	0.0119	0.0119
Logistic	0.7884 $\pm$	0.8444 $\pm$	0.7884 $\pm$	0.7799 $\pm$
Regression	0.0193	0.0163	0.0193	0.0209

Table 6 shows that XGBoost achieved the best mean accuracy, recall, and F1-score, while Random Forest remained a close competitor. Therefore, the advantage of XGBoost is not only numerical performance but also its ability to support SHAP-based interpretation of nonlinear feature effects. These models were selected because they are commonly used as comparative baselines in structured classification tasks, including RESTI-related machine learning studies [37].

The comparative results show that XGBoost achieved the highest mean accuracy, recall, and F1-score, while Random Forest achieved a slightly higher mean precision. The performance difference between XGBoost and Random Forest was modest, so the proposed model should not be interpreted as algorithmically superior in a universal sense. Instead, XGBoost was retained because it provided a strong balance between predictive performance, nonlinear interaction modeling, and post-hoc explanation through SHAP [18-20]. Figure 6 reports the ablation results used to assess whether engineered features and model selection improve prediction performance.

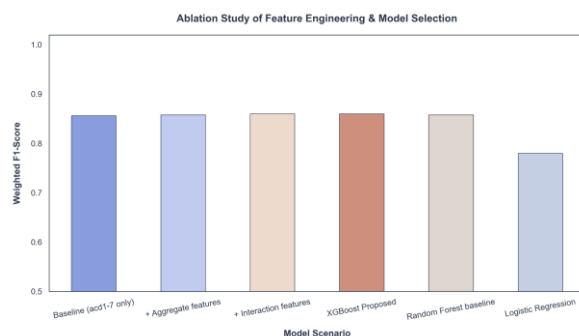


Figure 6. Ablation study of feature engineering and model selection.

The ablation analysis addresses the reviewer's concern that the contribution of the hybrid pipeline was previously under-tested. The baseline acd1-acd7 model already produced strong performance, indicating that academic stress indicators contain substantial predictive information. The addition of aggregate and interaction features resulted in small but consistent improvements, especially when paired with nonlinear tree-based models. Logistic Regression remained lower than the tree-based models, suggesting that nonlinear combinations among academic stressors contribute to prediction. Figure 7 evaluates the high-risk decision threshold because missed high-risk cases are the most consequential screening error.

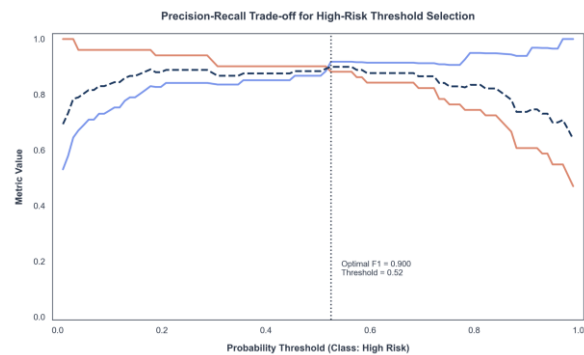


Figure 7. High-risk threshold trade-off analysis.

The threshold analysis clarifies the operational use of the early-warning framework. Because the most consequential error is missing a high-risk student, institutions may select a lower high-risk threshold to increase recall even if this increases false alerts. In the present internal analysis, a threshold of 0.52 provided a favorable high-risk F1-score. However, any threshold used in practice should be validated with counseling capacity, human review, and institutional policy before deployment. Figure 8 compares academic stress profiles between detected and missed high-risk students to illustrate the model failure pattern.



Figure 8. Feature profile comparison of detected and missed high-risk cases.

The high-risk error profile shows that correctly detected high-risk cases tended to have consistently high academic stress scores across multiple indicators, whereas missed high-risk cases displayed lower or more uneven academic stress profiles. This suggests that some high-risk students may not express uniformly high academic stress across all indicators, making them

harder to identify from academic stress features alone. This reinforces the need for human review and complementary professional assessment when the model is used for prioritization. Figure 9 assesses whether predicted probabilities are reasonably aligned with observed risk outcomes.

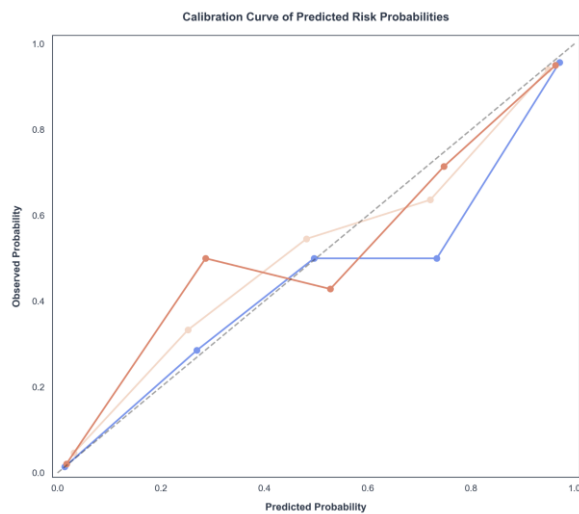


Figure 9. Calibration curve of predicted risk probabilities.

The calibration curve suggests that the model probabilities are informative but not perfectly calibrated across all probability ranges. Therefore, predicted probabilities should be used as prioritization scores rather than exact estimates of clinical risk.

The calibration curve indicates generally plausible probability behavior, particularly at higher predicted probabilities, but also shows deviations in the mid-probability range. These deviations mean that probability scores should not be interpreted as exact clinical risk probabilities. For practical deployment, the model should be recalibrated and externally tested before probability thresholds are used as institutional decision rules.

3.3 Subgroup Distribution and Bias Screening

Figure 10 presents the demographic distribution of risk classes to screen visible subgroup imbalance. The distribution plot shows that high-risk proportions vary moderately across study programs, faculties, and cohorts, while gender differences are relatively small. These descriptive differences justify subgroup monitoring but do not imply causal demographic effects.

The subgroup distribution was broadly comparable by gender, while study program, faculty, and cohort showed modest variation. These descriptive patterns should not be interpreted as causal demographic effects. Figure 11 reports subgroup performance to identify variation in accuracy, F1-score, high-risk recall, and false-negative rate.

The heatmap identifies variation in high-risk recall. Higher false-negative rates in some subgroups indicate

that deployment should include periodic fairness monitoring.

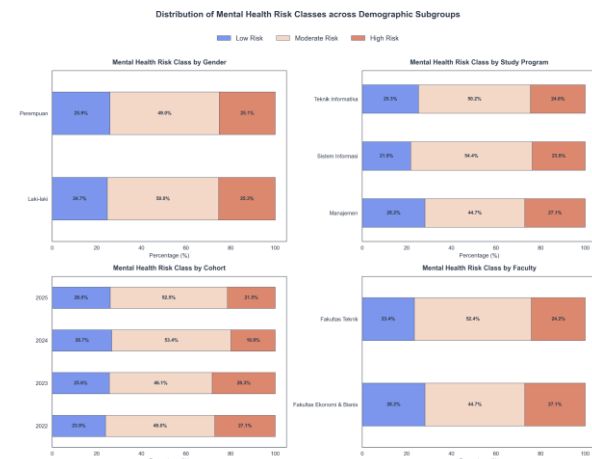


Figure 10. Risk class distribution across demographic subgroups.

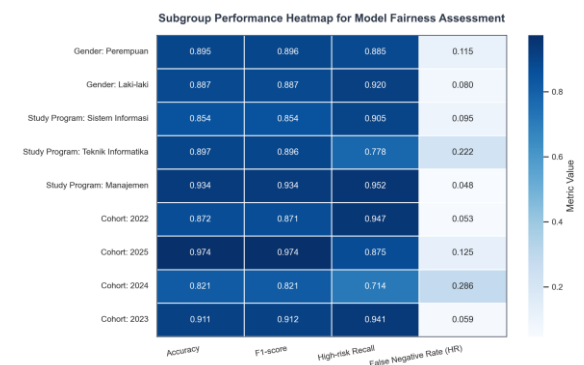


Figure 11. Subgroup performance heatmap for fairness screening.

Most subgroups achieved high accuracy and F1-score values. However, high-risk false-negative rates were higher in the 2024 cohort and Informatics Engineering subgroup. This does not establish discriminatory bias because subgroup sizes are smaller than the full test set, but it indicates where follow-up validation is needed. The 2024 cohort may still be adapting to university assessment norms and may report pressure less uniformly, while Informatics Engineering stress may appear through project-based workload, programming-debugging pressure, and deadline cycles that are not always captured by all seven academic stress items.

3.4 ROC, Precision-Recall, and Probability-Based Evaluation

Figure 12 evaluates one-vs-rest class discrimination for each risk category using ROC analysis. The ROC curves indicate strong separability for low- and high-risk classes, while moderate risk remains more difficult because it occupies a transitional position between the two extremes. ROC-AUC was included because it supports class-discrimination assessment beyond accuracy and is widely used for classifier audit and explanation [38].

The one-vs-rest ROC analysis confirms strong discrimination across classes. The AUC values were

approximately 0.99 for low risk, 0.95 for moderate risk, and 0.98 for high risk. The moderate-risk class remained the most difficult to separate because it lies between low- and high-risk profiles, which is consistent with the PCA visualization and adjacent-category misclassification pattern.

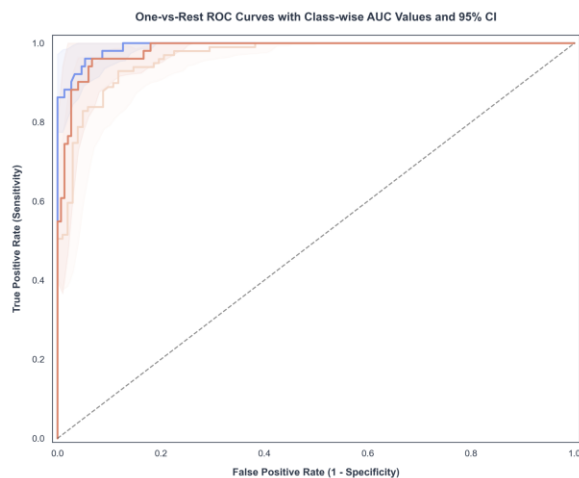


Figure 12. One-vs-rest ROC curves with class-wise AUC values.

Figure 13 evaluates precision-recall behavior across risk classes, which is important for screening settings with potential false alerts and missed cases.

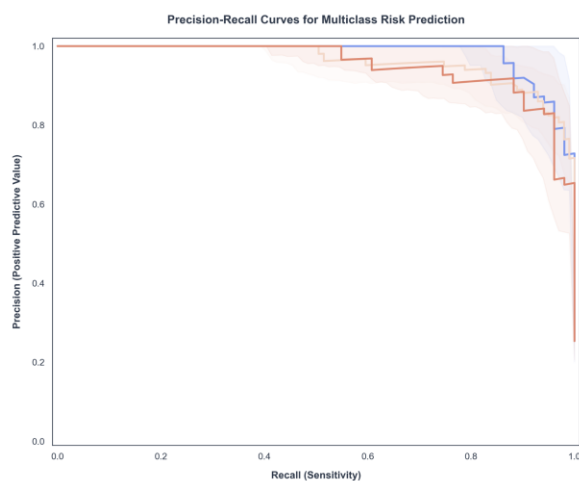


Figure 13. Precision-recall curves for multiclass risk prediction.

The precision-recall curves confirm that the model maintains strong precision at high recall ranges, which is relevant for screening use where missed high-risk students are more consequential than ordinary classification errors.

The precision-recall analysis further supports model performance under class-specific evaluation. This analysis is particularly relevant in screening settings because a high number of false alerts may burden counseling services, whereas low recall may delay support for students who need further assessment. The high-risk curve remained strong, but threshold selection should prioritize recall when the institutional objective is early support rather than strict classification accuracy.

3.5 SHAP Interpretation and Stability

Figure 14 shows the SHAP distribution for high-risk prediction to explain both feature importance and the direction of feature effects.

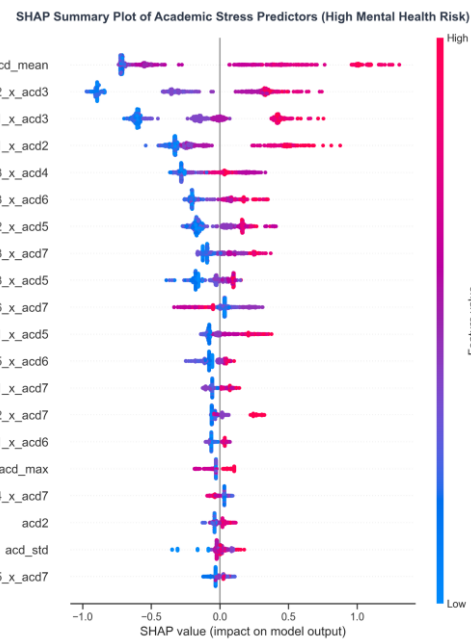


Figure 14. SHAP summary plot for high-risk prediction.

The beeswarm plot confirms that higher cumulative academic stress generally pushes the model toward high-risk predictions. It also shows that interaction features involving exam anxiety and learning-method adaptation contribute meaningfully to high-risk classification.

The SHAP beeswarm plot indicates that *acd_mean* was the most influential predictor for high-risk prediction. High cumulative academic stress values pushed predictions toward the high-risk class. Several interaction features also appeared among the dominant predictors, especially *acd2_x_acd3*, *acd1_x_acd3*, and *acd1_x_acd2*. These patterns indicate that parental expectation-related pressure, exam anxiety, and learning-method adaptation stress interact in shaping the model's risk prediction. Figure 15 evaluates whether top SHAP feature rankings remain consistent across cross-validation folds.

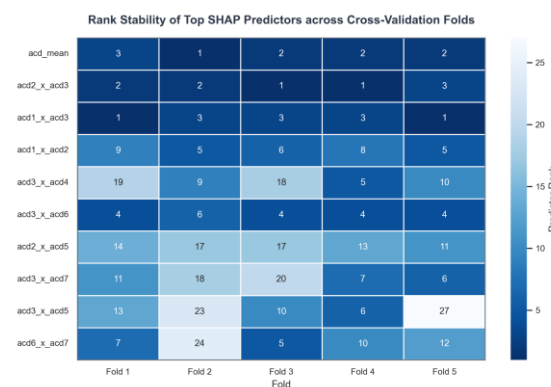


Figure 15. SHAP rank stability across cross-validation folds.

The rank-stability heatmap shows that the strongest predictors remain relatively consistent across folds, whereas lower-ranked interactions fluctuate more. The discussion therefore focuses on stable top predictors rather than overinterpreting every feature.

The SHAP rank stability analysis shows that `acd_mean`, `acd2_x_acd3`, and `acd1_x_acd3` consistently appeared among the top-ranked predictors across folds. Some lower-ranked interaction features varied substantially across folds, indicating that their importance may be more sample-specific. The interpretation therefore emphasizes the most stable predictors rather than every interaction feature. Figure 16 reports bootstrap-based SHAP stability to assess uncertainty around feature importance estimates.

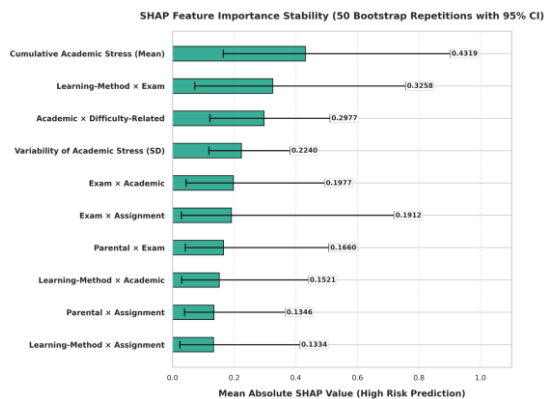


Figure 16. Bootstrap-based SHAP feature importance stability.

The bootstrap analysis adds uncertainty information to the SHAP interpretation. Cumulative academic stress remains the most reliable predictor, while wider intervals for several interactions indicate that their importance should be interpreted cautiously.

The bootstrap-based SHAP stability analysis further indicates that cumulative academic stress remained among the most consistently influential features. However, several interaction features had wider uncertainty ranges, which suggests that their attribution strength should be interpreted cautiously. These results address the reviewer's concern that SHAP explanations should not be based solely on a single training-test split.

Table 7. SHAP-based interpretation of stable influential academic stress features.

Important Feature	Meaning	Institutional Implication
<code>acd_mean</code>	Cumulative academic stress	Early mentoring for students with broad stress exposure
<code>acd2_x_acd3</code>	Parental expectation & Exam anxiety	Exam and grading support program; expectation counseling
<code>acd1_x_acd3</code>	Learning method & Exam anxiety	Lecturer communication and assessment transparency
<code>acd_std</code>	Variance in academic stressors	Targeted support for specific acute academic issues
<code>acd1</code>	Learning method pressure	Review teaching methodologies and workload

Table 7 translates the stable SHAP predictors into practical meanings and possible institutional implications. Table 7 converts the stable SHAP predictors into institutional meaning. The table indicates that cumulative stress and combined pressures involving parental expectations, exam anxiety, and learning-method adaptation are the most actionable signals for early support.

3.6 Practical, Ethical, and Deployment Implications

The revised analysis clarifies that the framework is an assistive screening tool for prioritizing academic mentoring or counseling referral. High-risk predictions should trigger human review by authorized personnel, not permanent labeling or clinical diagnosis [22], [23].

The Streamlit-based prototype demonstrates how the framework can be operationalized in educational settings. Users can input academic stress indicators and obtain automated risk classifications with SHAP-based explanations. However, deployment requires strict governance: informed data use, restricted access, logging of model outputs, clear referral protocols, periodic recalibration, and monitoring of subgroup performance. Because the current model was trained on 1,002 records from UMK, the Streamlit prototype should be implemented only after local retraining and recalibration with institutional data; otherwise, local teaching culture, assessment practices, counseling procedures, and student response patterns may shift feature-risk relationships. These safeguards are necessary because the model is based on a single-site dataset and single-item academic stress indicators.

3.7 Limitations and Future Work

This study has several limitations. First, the dataset was collected from a single institution, so local curriculum, teaching culture, assessment practices, student-support procedures, and data-collection timing may influence the model. The trained model may therefore overfit to the institutional characteristics of Universitas Muria Kudus and should not be transferred directly to other universities. Institutions outside UMK should treat the current system as a transferable workflow rather than a ready-to-use trained model; local retraining, recalibration, and external validation are required before implementing the Streamlit prototype. Second, the risk labels were derived from K-Means clustering and therefore depend on DASS-21 score construction, normalization, and cluster interpretation.

Although sensitivity and stability analyses were added, cluster-derived labels remain empirical categories rather than clinical diagnoses.

Academic stress dimension was operationalized using a single-item Likert indicator. Single-item measures are practical for early screening but cannot fully capture the reliability and construct validity of multidimensional academic stress constructs. Future research should use validated multi-item academic stress scales and report psychometric reliability. Fourth, the validation remains

internal. Although hold-out testing, cross-validation, subgroup analysis, calibration, threshold analysis, and SHAP stability checks improve robustness, they do not replace external, temporal, or prospective validation. Fifth, SHAP explanations should be interpreted as model attribution, not causal evidence.

Future work should evaluate the model in multiple universities, test temporal stability across academic semesters, incorporate validated multi-item stress instruments, compare alternative clustering algorithms, and examine whether human-in-the-loop review improves support outcomes. Deployment studies should also evaluate counseling workload, false-alert management, student consent, and the fairness of referral decisions.

4. Conclusions

This study developed and revised an explainable early-warning framework for mapping student mental health risk from academic stress indicators. K-Means clustering was used to construct low-, moderate-, and high-risk categories from DASS-21 subscale profiles, while XGBoost predicted the cluster-derived labels from academic stress indicators and engineered features. On the 201-record hold-out test set, the model produced weighted precision, recall, and F1-score values of 0.8907, 0.8905, and 0.8906, respectively.

The revised analysis strengthens the manuscript by adding ablation analysis, K-Means sensitivity and stability testing, annotated error analysis, threshold evaluation, calibration diagnostics, demographic subgroup analysis, SHAP beeswarm interpretation, SHAP rank stability, and bootstrap-based SHAP stability. These additions address the reviewers' concerns about methodological transparency, robustness, subgroup performance, and interpretability.

The findings indicate that cumulative academic stress and interactions involving parental expectation, exam anxiety, and learning-method adaptation are important signals for predicting high-risk profiles. Nevertheless, the model is intended only as non-clinical decision support for early institutional prioritization and counseling referral. Because the study uses a single-site dataset and single-item academic stress indicators, external validation and psychometric strengthening are required before broader deployment. Any institution intending to use the Streamlit prototype must retrain and recalibrate the model on local student data before implementation.

The framework was also translated into a Streamlit-based prototype, demonstrating that the proposed model can be operationalized as an interactive and explainable early-warning interface for authorized academic support stakeholders. In practical use, the prototype should function as a locally retrained decision-support interface rather than an automatically transferable diagnostic system.

References

- [1] S. K. Lipson *et al.*, "Trends in college student mental health and help-seeking by race/ethnicity: Findings from the national healthy minds study, 2013–2021," *J. Affect. Disord.*, vol. 306, pp. 138–147, 2022, doi: <https://doi.org/10.1016/j.jad.2022.03.038>.
- [2] Y. Deng, J. Cherian, N. Un, N. Khan, and K. Kumari, "Family and Academic Stress and Their Impact on Students' Depression Level and Academic Performance," *Front. Psychiatry*, vol. 13, no. June, pp. 1–13, 2022, doi: [10.3389/fpsy.2022.869337](https://doi.org/10.3389/fpsy.2022.869337).
- [3] G. Barbayannis, M. Bandari, X. Zheng, H. Baquerizo, K. W. Pecor, and X. Ming, "Academic Stress and Mental Well-Being in College Students: Correlations, Affected Groups, and COVID-19," *Front. Psychol.*, vol. 13, no. May, pp. 1–10, 2022, doi: [10.3389/fpsyg.2022.886344](https://doi.org/10.3389/fpsyg.2022.886344).
- [4] D. T. H. Le, R. C. K. Chan, J. Hao, and R. B. Wilkinson, "The Mental Health of Australian, Chinese and Vietnamese University Students: A Cross-Cultural Comparison of the DASS-21," *Int. J. Psychol.*, vol. 60, no. 4, 2025, doi: [10.1002/ijop.70057](https://doi.org/10.1002/ijop.70057).
- [5] M. A. Castillo-Díaz and C. A. H. Periañez, "Dimensionality, reliability, invariance, and item analysis of the depression, anxiety, and stress scale-21 (DASS-21) in Honduran and Colombian university students," *BMC Psychol.*, vol. 13, no. 1, 2025, doi: [10.1186/s40359-025-03742-5](https://doi.org/10.1186/s40359-025-03742-5).
- [6] M. E. Hasan *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. Heal. Popul. Nutr.*, vol. 44, no. 1, 2025, doi: [10.1186/s41043-025-01095-8](https://doi.org/10.1186/s41043-025-01095-8).
- [7] J. Chen, "Student Mental Health Risk Prediction Based on Apriori Algorithm in the Context of Big Data," in *2024 International Conference on Electrical Drives, Power Electronics & Engineering (EDPEE)*, 2024, pp. 615–618. doi: [10.1109/EDPEE61724.2024.00120](https://doi.org/10.1109/EDPEE61724.2024.00120).
- [8] M. F. Chan, A. Al Harrasi, H. Al Sinawi, and M. Al Alawi, "Sociodemographics and clinical profiling of students seeking mental health services : a cluster analysis in a public university in Oman," *Learn. Teach. High. Educ. Gulf Perspect.*, 2026, doi: [10.1108/LTHE-07-2025-0048](https://doi.org/10.1108/LTHE-07-2025-0048).
- [9] F. N. Al-Wesabi, H. Alsolai, A. M. Hilal, M. A. Hamza, M. Al Duhayyim, and N. Negm, "Machine learning based depression, anxiety, and stress predictive model during COVID-19 crisis," *Comput. Mater. Contin.*, vol. 70, no. 3, pp. 5803–5820, 2022, doi: [10.32604/cmc.2022.021195](https://doi.org/10.32604/cmc.2022.021195).
- [10] D. Duan, "Using a predictive model based on fuzzy logic methods to test the mental health of students," *Discov. Comput.*, vol. 29, no. 1, 2026, doi: [10.1007/s10791-026-09962-6](https://doi.org/10.1007/s10791-026-09962-6).
- [11] K. Lamba, S. Rani, and M. Shabaz, "Explainable machine learning for mental health prediction from social media behavior: a nested cross-validation study with {SHAP} and {LIME} interpretability," *Discov. Ment. Heal.*, vol. 6, no. 1, 2026, doi: [10.1007/s44192-026-00373-z](https://doi.org/10.1007/s44192-026-00373-z).
- [12] Y. Wang and Y. Tong, "A Machine Learning Approach to Explore Perceived Stress in College Students Post {COVID-19} Pandemic," *Int. J. Inf. Syst. Model. Des.*, vol. 16, no. 1, pp. 1–21, 2025, doi: [10.4018/IJISMD.391327](https://doi.org/10.4018/IJISMD.391327).
- [13] A. Marika, "Identifying Profiles Of Psychological Distress , Psychological Flexibility , And Mindfulness Among Finnish University Students Participating In A Blended Intervention," *Univ. JYVÄSKYLÄ Dep. Psychol.*, no. May, 2025.
- [14] K. N. Singh, H. P. Singh, M. Panjwani, and S. Dahima, "Mental Health Detection and Prediction through Machine Learning Technology," in *Brain Informatics Technology*, John Wiley & Sons, Ltd, 2025, ch. 16, pp. 403–423. doi: <https://doi.org/10.1002/9781394345625.ch16>.
- [15] S. Risal, F. Apriyadi, A. Sumardin, A. D. Achmad, and A. N. Puteri, "Enhancing Stroke Prediction with Logistic Regression and Support Vector Machine Using Oversampling Techniques," *J. RESTI (Rekayasa Sist. dan Teknol. Informatika)*, vol. 9, no. 3, pp. 646–658, 2025, doi: [10.29207/resti.v9i3.6431](https://doi.org/10.29207/resti.v9i3.6431).
- [16] H. D. Putra, L. Khairani, and D. Hastari, "Comparison of Naive Bayes Classifier and Support Vector Machine Algorithms for

- Classifying Student Mental Health Data,” in *SENTIMAS: Seminar Nasional Penelitian dan Pengabdian Masyarakat*, 2023, pp. 120–125. [Online]. Available: <https://journal.irpi.or.id/index.php/sentimas/article/view/586/347>
- [17] S. Kapoor and A. Narayanan, “Leakage and the reproducibility crisis in machine-learning-based science,” *Patterns*, vol. 4, no. 9, p. 100804, 2023, doi: [10.1016/j.patter.2023.100804](https://doi.org/10.1016/j.patter.2023.100804).
- [18] H. Guo, H. Zhang, and T. Erdenebold, “An interpretable cross-regional prediction study of university students’ mental health based on XGBoost and SHAP,” *IEEE Access*, vol. 14, no. April, pp. 48025–48039, 2026, doi: [10.1109/ACCESS.2026.3676575](https://doi.org/10.1109/ACCESS.2026.3676575).
- [19] R. Schwartz-Ziv and A. Armon, “Tabular data: Deep learning is not all you need,” *Inf. Fusion*, vol. 81, pp. 84–90, 2022, doi: <https://doi.org/10.1016/j.inffus.2021.11.011>.
- [20] S. Ali et al., “Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence,” *Inf. Fusion*, vol. 99, p. 101805, 2023, doi: <https://doi.org/10.1016/j.inffus.2023.101805>.
- [21] P. Rajpurkar, E. Chen, O. Banerjee, and E. J. Topol, “AI in health and medicine,” *Nat. Med.*, vol. 28, no. 1, pp. 31–38, 2022, doi: [10.1038/s41591-021-01614-0](https://doi.org/10.1038/s41591-021-01614-0).
- [22] B. Vasey et al., “Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI,” *Nat. Med.*, vol. 28, no. 5, pp. 924–933, 2022, doi: [10.1038/s41591-022-01772-9](https://doi.org/10.1038/s41591-022-01772-9).
- [23] N. Naik et al., “Legal and Ethical Consideration in Artificial Intelligence in Healthcare: Who Takes Responsibility?,” *Front. Surg.*, vol. 9, no. March, pp. 1–6, 2022, doi: [10.3389/fsurg.2022.862322](https://doi.org/10.3389/fsurg.2022.862322).
- [24] P. Córdova Olivera, P. Gasser Gordillo, H. Naranjo Mejía, I. La Fuente Taborga, A. Grajeda Chacón, and A. Sanjinés Unzueta, “Academic stress as a predictor of mental health in university students,” *Cogent Educ.*, vol. 10, no. 2, 2023, doi: [10.1080/2331186X.2023.2232686](https://doi.org/10.1080/2331186X.2023.2232686).
- [25] R. R. Bojja, “A Comparative Study of Supervised and Unsupervised Learning Approaches,” *Imrjr*, vol. 2, no. 4, pp. 80–88, 2025, doi: [10.17148/imrjr.2025.020411](https://doi.org/10.17148/imrjr.2025.020411).
- [26] I. H. Sarker, “Machine Learning: Algorithms, Real-World Applications and Research Directions,” *SN Comput. Sci.*, vol. 2, no. 3, p. 160, 2021, doi: [10.1007/s42979-021-00592-x](https://doi.org/10.1007/s42979-021-00592-x).
- [27] C. Schröer, F. Kruse, and J. M. Gómez, “A Systematic Literature Review on Applying CRISP-DM Process Model,” *Procedia Comput. Sci.*, vol. 181, pp. 526–534, 2021, doi: [10.1016/j.procs.2021.01.199](https://doi.org/10.1016/j.procs.2021.01.199).
- [28] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, “K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data,” *Inf. Sci. (Ny)*, vol. 622, pp. 178–210, 2023, doi: <https://doi.org/10.1016/j.ins.2022.11.139>.
- [29] W. Chen, K. Yang, Z. Yu, Y. Shi, and C. L. P. Chen, *A survey on imbalanced learning: latest research, applications and future directions*, vol. 57, no. 6, 2024, doi: [10.1007/s10462-024-10759-6](https://doi.org/10.1007/s10462-024-10759-6).
- [30] S. Rahmawati, A. Wibowo, and A. F. N. Masruriyah, “Improving Diabetes Prediction Accuracy in Indonesia: A Comparative Analysis of {SVM}, Logistic Regression, and Naive Bayes with {SMOTE} and {ADASYN},” *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 8, no. 5, pp. 607–614, 2024, doi: [10.29207/resti.v8i5.5980](https://doi.org/10.29207/resti.v8i5.5980).
- [31] L. Longo et al., “Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions,” *Inf. Fusion*, vol. 106, no. February, p. 102301, 2024, doi: [10.1016/j.inffus.2024.102301](https://doi.org/10.1016/j.inffus.2024.102301).
- [32] N. L. Wibowo, M. A. Soeleman, and A. Z. Fanani, “Antlion Optimizer Algorithm Modification for Initial Centroid Determination in K-means Algorithm,” *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 7, no. 4, pp. 870–883, 2023, doi: [10.29207/resti.v7i4.4997](https://doi.org/10.29207/resti.v7i4.4997).
- [33] A. E. Ezugwu et al., “A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects,” *Eng. Appl. Artif. Intell.*, vol. 110, p. 104743, 2022, doi: <https://doi.org/10.1016/j.engappai.2022.104743>.
- [34] K. K. Pandey and D. Shukla, “Min–max kurtosis mean distance based k-means initial centroid initialization method for big genomic data clustering,” *Evol. Intell.*, vol. 16, no. 3, pp. 1055–1076, 2023, doi: [10.1007/s12065-022-00720-3](https://doi.org/10.1007/s12065-022-00720-3).
- [35] R. Wardhani, N. Nafiiyah, and M. Kemal Farouq, “K-Means Clustering of the Mental Health of Engineering Students at the Universitas Islam Lamongan,” in *2023 IEEE International Conference on Communication, Networks and Satellite (COMNETSAT)*, 2023, pp. 181–186, doi: [10.1109/COMNETSAT59769.2023.10420590](https://doi.org/10.1109/COMNETSAT59769.2023.10420590).
- [36] G. T. Reddy et al., “Analysis of Dimensionality Reduction Techniques on Big Data,” *IEEE Access*, vol. 8, pp. 54776–54788, 2020, doi: [10.1109/ACCESS.2020.2980942](https://doi.org/10.1109/ACCESS.2020.2980942).
- [37] S. Liang, “Comparative Analysis of SVM, XGBoost and Neural Network on Hate Speech Classification,” *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 5, pp. 896–903, 2021, doi: [10.29207/resti.v5i5.3506](https://doi.org/10.29207/resti.v5i5.3506).
- [38] A. M. Carrington et al., “Deep ROC Analysis and AUC as Balanced Average Accuracy, for Improved Classifier Selection, Audit and Explanation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 329–341, 2023, doi: [10.1109/TPAMI.2022.3145392](https://doi.org/10.1109/TPAMI.2022.3145392).