

Comparative Machine Learning Approaches for Classifying Student Stress Types

Chrisensia Abigail Gani, Clarice Harijanto, Christopher Andreas*

School of Information Technology, Universitas Ciputra, Surabaya, Indonesia
Email: ¹cabigail03@student.ciputra.ac.id, ²charijanto01@student.ciputra.ac.id, ^{3,*}christopher.andreas@ciputra.ac.id
Correspondence Author Email: christopher.andreas@ciputra.ac.id

Received: 17-09-2026, Revised: 06-10-2026, Accepted: 07-10-2026, Published: 19-10-2026

Abstract—Student stress is a major mental health concern that may affect academic performance, psychological well-being, and quality of life. Early identification of stress types is important because different stress categories may require different intervention strategies. With advances in data-driven technologies, machine learning has emerged as a promising approach for identifying patterns in mental health data and supporting automated assessment. This study aims to classify student stress types using the Stress Indicators Dataset for Mental Health Classification, which contains 2,000 student responses. The dataset includes demographic information, psychological symptoms, academic challenges, physical health indicators, and social factors associated with stress. Before model development, preprocessing procedures included data cleaning, duplicate removal, filtering extreme age values, and exploratory data analysis (EDA). Feature engineering was also performed by creating three composite variables, namely symptom score, academic stress score, and social_stress_score, to represent broader stress-related dimensions. Three classification algorithms, Logistic Regression, K-Nearest Neighbors (KNN), and Gaussian Naive Bayes, were evaluated using stratified five-fold cross-validation, with Accuracy, Balanced Accuracy, Macro Precision, Macro Recall, and Macro F1-score. Results indicate that Logistic Regression achieved the highest overall performance, with $97.59\% \pm 1.09\%$ accuracy and $88.15\% \pm 5.18\%$ Macro F1-score. Gaussian Naive Bayes with engineered features achieved the highest Balanced Accuracy of $91.45\% \pm 5.33\%$. Feature engineering slightly improved KNN performance but did not improve Logistic Regression or Gaussian Naive Bayes accuracy. Given the substantial class imbalance, these results should be interpreted alongside class-specific metrics and do not establish clinical applicability or broad generalizability across other student populations.

Keywords: Feature Engineering; Machine Learning; Mental Health Assessment; Stress Classification; Student Stress

1. INTRODUCTION

Mental health has become an increasingly urgent global concern, with recent evidence indicating that a substantial proportion of students experience significant psychological distress during their studies. A 2025 Gallup survey found that approximately one in three U.S. postsecondary students had considered withdrawing from their academic program within the past six months, with emotional stress and personal mental health concerns cited as the most common reasons [1]. This happens not only in the U.S. but internationally as well. A World Health Organization (WHO) supported study conducted across nine countries found that among more than 20,000 first-year university students, 93.7% experienced stress in at least one major life domain, including academic, financial, social, and health-related areas, with elevated stress levels associated with an increased risk of depression and anxiety [2]. When left unidentified and unaddressed, prolonged student stress can develop into more severe psychological conditions, including clinical anxiety, depression, and academic burnout, ultimately compromising both psychological wellbeing and academic performance [3].

Stress is not a consistent experience for students. Numerous factors, including demographics, psychological symptoms, academic demands, physical health, and social environment, influence it. Recent research that combines traditional statistical methods with machine learning has demonstrated the multidimensional correlates of psychological stress in nationally representative populations [4]. Selye suggested that stress can be identified as either eustress, a constructive and motivating form, or distress, a negative and destructive form, underscoring that different stress types may warrant different intervention strategies [5]. Therefore, categorising the particular kind of stress a student encounters provides educational institutions with far more useful information than binary detection of stress presence alone, indicating whether peer support, academic advising, or counselling is the better course of action. Conventional stress assessment typically relies on psychological questionnaires interpreted manually by trained professionals. Even though it is effective, this approach requires considerable time and resources, particularly when applied to large student populations, limiting its scalability within higher education institutions [6]. Machine learning has emerged as a promising alternative, offering the capacity to automatically identify complex patterns within large-scale survey data and support more efficient, systematic mental health assessment [7]. However, for institutional screening applications, predictive accuracy should not be considered the sole criterion for model selection. Models used to support student interventions should also provide sufficiently transparent decision processes, manageable computational requirements, and reproducible results, particularly when their predictions may inform subsequent academic advising or counselling decisions. Thus, simpler and more interpretable classifiers can represent a practically relevant alternative to highly complex models, even when the latter may provide competitive predictive performance.

Several recent studies have applied machine learning and statistical approaches to student and academic stress. Hives et al. combined traditional statistical modeling with machine learning techniques to identify multidimensional correlates of psychological stress using a nationally representative Canadian sample, demonstrating that stress is better explained through multiple interacting factors than through single-variable models, however, the study focused on identifying correlates rather than developing a deployable classification framework for stress type prediction [4]. Anand



et al. developed an ensemble learning framework using student survey data on sleep duration, screen time, and academic workload, achieving 93.48% accuracy in classifying stress into three severity levels, with ensemble methods (Random Forest, AdaBoost, Gradient Boosting) consistently outperforming individual classifiers. Nonetheless, the study addressed stress severity rather than stress type, and its reliance on methods limits interpretability for non-technical parties such as academic counselors [8]. Schaab et al. conducted a systematic review of machine learning applications for detecting depression, anxiety, and stress among undergraduate students, finding considerable methodological heterogeneity across studies and noting that most focused on binary classification (stress present or absent) rather than multi-class stress typing, with few studies systematically benchmarking simple, interpretable classifiers against one another [9]. Daza et al. similarly reviewed machine learning techniques for predicting anxiety and stress in college students and concluded that no clear consensus exists regarding which algorithm performs optimally, as performance is highly dependent on dataset characteristics and stress conceptualization [10]. Ruiz-Camacho and Gozalo observed the predictive role of academic stressors and sociodemographic variables across five stress response dimensions using hierarchical regression, finding that academic stressors accounted for substantially more variance than sociodemographic factors, while informative, this study relied on regression-based prediction rather than classification, and did not address multi-class stress type identification through machine learning [11].

These studies indicate that, although machine learning has been increasingly applied to student stress research, existing studies largely focus on binary stress detection, rely on complex black-box models with limited interpretability, or examine associations rather than developing practical classification systems. However, systematic comparisons of simple and interpretable models, particularly Logistic Regression (LR), K-Nearest Neighbors (KNN), and Gaussian Naive Bayes (GNB), for classifying student stress types remain limited. The selection of these three models is also motivated by a decision-oriented consideration of model complexity. LR provides a transparent linear decision mechanism and readily interpretable coefficients, KNN offers a non-parametric distance-based perspective without imposing a predefined functional relationship, and GNB provides a computationally efficient probabilistic approach based on conditional feature distributions. Evaluating these complementary model structures allows the study to determine whether more complex predictive mechanisms are actually necessary for this classification task, rather than assuming that model complexity inherently leads to better practical decisions [12-15].

Furthermore, the use of feature engineering to improve predictive performance while preserving interpretability for practical institutional applications has received insufficient attention. This consideration is particularly important in educational settings where predictions may be used as decision-support information rather than as autonomous clinical diagnoses. From this perspective, the preferred model should achieve an appropriate balance between predictive effectiveness, interpretability, computational simplicity, and potential deployment feasibility. To address these gaps, this study evaluates three classification algorithms, LR, KNN, and GNB, representing linear, distance-based, and probabilistic approaches, respectively, for classifying student stress into three types: Distress, Eustress, and Other/Mixed Stress, using the Stress Indicators Dataset for Mental Health Classification [16]. Rather than presuming that advanced algorithms will necessarily provide superior results, the study treats model selection as an empirical and decision-oriented comparison in which predictive performance is evaluated alongside interpretability and practical feasibility. In addition to the baseline models, three composite features, namely `symptom_score`, `academic_stress_score`, and `social_stress_score`, are constructed to capture broader stress-related dimensions and improve model interpretability. Model performance is assessed using several metrics, including Accuracy, Precision, Recall, F1-Score, and Confusion Matrix analysis. The main contributions of this study are threefold. First, it systematically compares three relatively simple machine learning models for multi-class student stress classification. Second, it develops interpretable composite features from raw survey responses to represent broader dimensions of student stress. Third, it examines the balance between predictive performance and model interpretability, providing a decision-oriented assessment of whether lightweight and transparent classifiers can offer a sufficiently effective alternative to more complex machine learning approaches for educational institutions with limited technical resources seeking to implement automated student stress screening tools.

2. RESEARCH METHODOLOGY

2.1 Research Stages

The research object in this study is the Stress Indicators Dataset for Mental Health Classification [12], which represents questionnaire-based responses related to student stress. The study analyzed the Stress Indicators Dataset for Mental Health Classification [16]. The dataset consists of 2,000 student responses collected via a questionnaire using Likert scale, ranging from 1 ("Not at all") to 5 ("Extremely"). The survey items cover demographic characteristics, psychological symptoms, physical health indicators, academic stressors, and social environment factors. The target variable, `stress_type`, contains three distinct classes: Distress (0), Eustress (1), and Other/Mixed Stress (2). The three-class structure makes the dataset suitable for evaluating multi-class classification approaches rather than binary stress detection alone. The class distribution was examined during exploratory data analysis to assess potential class imbalance and to determine its implications for model evaluation. Figure 1 summarizes the overall research workflow described above, comprising six main stages: data acquisition, preprocessing and exploratory data analysis, feature engineering, data partitioning and scaling, model training, and model evaluation.



The predictor variables evaluated in this study include gender, age, stress experience, heartbeat palpitations, anxiety_tension, sleep_problems, restlessness, headaches, irritability, concentration_problems, sadness low mood, health issues, loneliness isolation, academic overload, peer competition, relationship stress, professor difficulties, work environment, lack relaxation time, home environment, low academic confidence, subject confidence, academic conflicts, class attendance, and weight changes. The analytical workflow began by importing essential Python libraries for data manipulation, visualization, modeling, and evaluation, specifically Pandas, NumPy, Matplotlib, Seaborn, and Scikit-learn. Once loaded into the environment, the dataset was inspected to understand its underlying structure and feature types. Data preprocessing was then conducted to clean the data before model training. Missing values were checked, duplicate records were removed to prevent data redundancy, and extreme age values falling outside the target student demographic were filtered out. Of the 2,000 raw responses, 1,191 (59.6%) were exact duplicates and were removed, leaving 809 unique responses. A further 22 respondents older than 25 years were excluded, yielding a final analytical dataset of 787 responses (Distress = 27, Eustress = 719, Other/Mixed = 41). Exploratory data analysis was then performed on the cleaned dataset to identify key distributions and trends across stress indicators. The preprocessing stage also included an audit of the predictor variables to ensure that the target variable, stress_type, was not directly included as a predictor or used to construct any predictor feature.

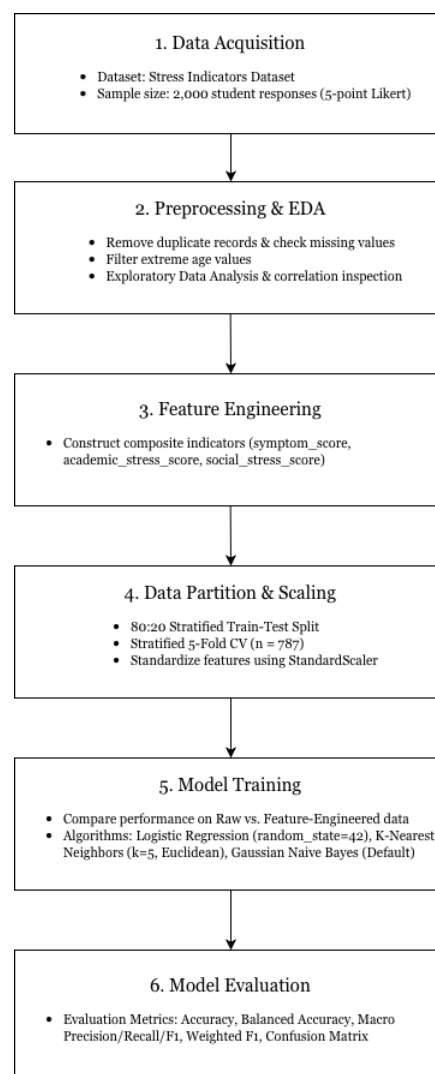


Figure 1. Overview of the research methodology.

Building on these insights, feature engineering was applied to construct three composite variables: symptom_score, academic_stress_score, and social_stress_score. These composite scores were created by aggregating related survey items to capture broader dimensions of student stress. Predictor variables (X) were separated from the target variable (y), and experiments were designed to compare model performance on both the raw feature set and the dataset enriched with composite features. The composite features were constructed exclusively from the corresponding questionnaire predictor variables and did not use stress_type or any information derived from other observations. This procedure was used to prevent target leakage during feature construction. The composite features were computed row-wise from each respondent's own questionnaire items, without using the target variable or any other respondent's data, so they could be constructed before data splitting without introducing leakage. Models were initially evaluated using a

stratified 80:20 train-test split. However, the resulting test set contained only small number of observations from the minority classes, limiting the reliability of class-specific performance estimates. Therefore, stratified five-fold cross-validation was subsequently performed on the full cleaned dataset and used as the primary evaluation procedure. In each iteration, four folds were used for training and one fold for validation, and the procedure was repeated until every observation had been used once for validation. Mean and standard deviation (SD) across the five folds were calculated to assess predictive performance and stability. The initial 80:20 split was retained only as a preliminary hold-out assessment, and its results were not used as the primary basis for model comparison. To prevent scaling leakage, StandardScaler was embedded in a pipeline and fitted only on the training portion of each fold before being applied to the corresponding validation fold. The same stratified folds (random_state = 42) were used for all models and both feature sets to ensure a consistent comparison.

A confusion matrix visualized classification performance across stress categories. Accuracy measured overall correctness, while Precision and Recall assessed prediction reliability and class detection. F1-Score provided a balanced measure of Precision and Recall. Weighted averaging was applied to Precision, Recall, and F1-Score to account for unequal class sizes. In addition, macro-averaged Precision, Recall, and F1-Score were considered to provide a more balanced assessment across classes, particularly in the presence of class imbalance. The confusion matrices were further examined to identify systematic misclassification patterns between Distress, Eustress, and Other/Mixed Stress.

2.2 Comparative Modeling Approach

This study employs three classification algorithms representing distinct learning paradigms: linear, distance-based, and probabilistic. LR is a supervised learning method that estimates class probabilities using a logistic function applied to a linear combination of predictor variables. It is widely adopted for classification due to its interpretability, computational efficiency, and effectiveness for linearly separable data [12]. KNN is a non-parametric, instance-based method that assigns a class based on the majority class among its k nearest observations in the feature space, commonly determined using Euclidean distance. Unlike LR, KNN makes no assumptions about the functional relationship between predictors and outcomes, allowing it to capture non-linear patterns, although its performance is sensitive to feature scaling and the selected value of k [13-14]. GNB is a probabilistic classifier based on Bayes' Theorem that assumes conditional independence among predictors and models continuous features using a Gaussian distribution within each class. Despite this simplifying assumption, it can achieve competitive classification performance while offering computational efficiency and interpretability [15]. These algorithms were selected based on their distinct theoretical foundations, relatively low computational requirements, and high interpretability, which are relevant for practical deployment in institutional settings with limited technical infrastructure. More specifically, the three algorithms provide complementary decision mechanisms: LR provides a transparent linear decision boundary and interpretable class probabilities, KNN captures local and potentially non-linear patterns without imposing a predefined functional relationship, and GNB provides a computationally efficient probabilistic approach. This combination enables an empirical comparison of predictive performance against practical considerations such as interpretability, computational efficiency, and deployment feasibility. Accordingly, the study does not assume that simpler algorithms are inherently superior to advanced methods; rather, it examines whether additional model complexity is necessary to obtain sufficiently effective predictions for the evaluated classification task.

All algorithms were implemented using the Scikit-learn framework. To ensure a fair and direct comparison, all models were trained using their default configurations, all models were implemented using standard Scikit-learn configurations without extensive hyperparameter optimization, with minor adjustments applied where necessary (see Table 1). Specifically, LR was set with random_state = 42 to guarantee reproducible results, KNN was configured with n_neighbors = 5 and metric = "euclidean", and GNB was executed using its default parameter configuration. The same cross-validation folds, feature-processing procedure, and evaluation metrics were applied consistently to all three algorithms and to both the original and feature-engineered feature sets. This common experimental setting was intended to ensure that observed performance differences could be attributed primarily to the classification approach and feature representation rather than differences in data partitioning or preprocessing. Prior to model training, feature standardization was performed using StandardScaler. This standardization step was essential because linear and distance-based models such as LR and KNN are sensitive to differences in feature magnitudes [17]. Within cross-validation, standardization was performed separately within each training fold to avoid information leakage. The identical scaling workflow was applied to both the original dataset and the feature-engineered dataset.

Table 1. Hyperparameter Configuration for Each Classification Algorithm

Algorithm	Key Parameter(s)	Value
Logistic Regression (LR)	random_state	42
K-Nearest Neighbors (KNN)	n_neighbors; metric	5; Euclidean
Gaussian Naive Bayes (GNB)	Default configuration	—

3. RESULT AND DISCUSSION

3.1 Exploratory Data Analysis

Prior to model evaluation, explanatory data analysis was conducted to examine the characteristics of the cleaned dataset and identify patterns related to student stress indicators. Figure 2(a) presents the age distribution before preprocessing. Most respondents were concentrated between 18 and 22 years old, indicating that the dataset primarily represents young adults. However, several extreme age values were observed outside the typical student age range, suggesting the presence of potential outliers or data entry inconsistencies.

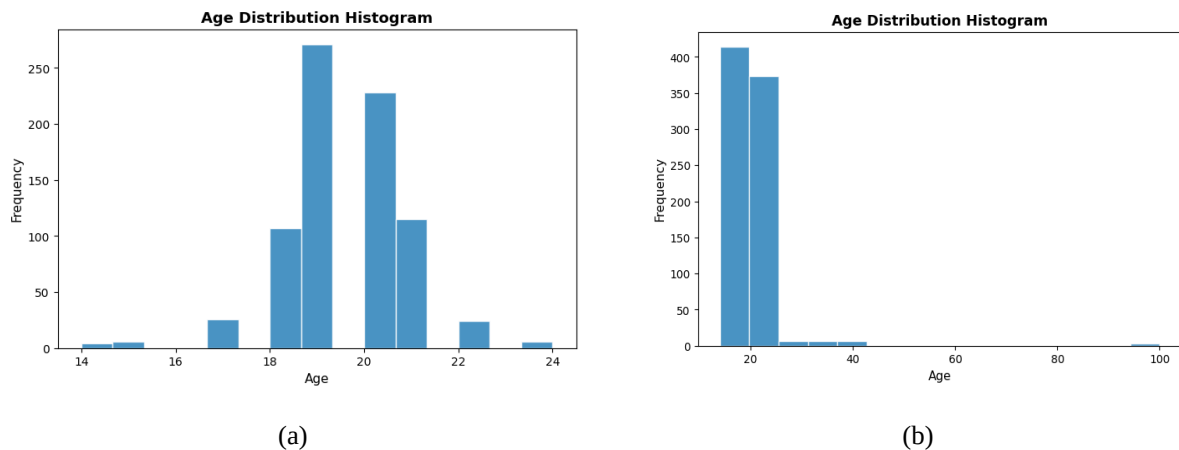


Figure 2. Age distribution (a) before preprocessing and (b) after preprocessing and age filtering.

After preprocessing and age filtering, the age distribution shown in Figure 2(b) became more concentrated within the 15–25 age range. This adjustment improved the representativeness of the dataset by focusing on the target student population while reducing the influence of extreme observations that could negatively affect model performance. Figure 3 illustrates the correlation heatmap of the engineered features and the target variable. The results indicate that all three engineered features were associated with stress classification. Among them, *symptom_score* exhibited the strongest relationship with *stress_type* ($r = -0.51$), followed by *academic_stress_score* ($r = -0.40$) and *social_stress_score* ($r = -0.33$). These findings suggest that psychological and physical symptoms play a particularly important role in distinguishing different stress categories. In addition, the correlations among the engineered features were generally weak to moderate, indicating that each feature captures a distinct dimension of student stress.

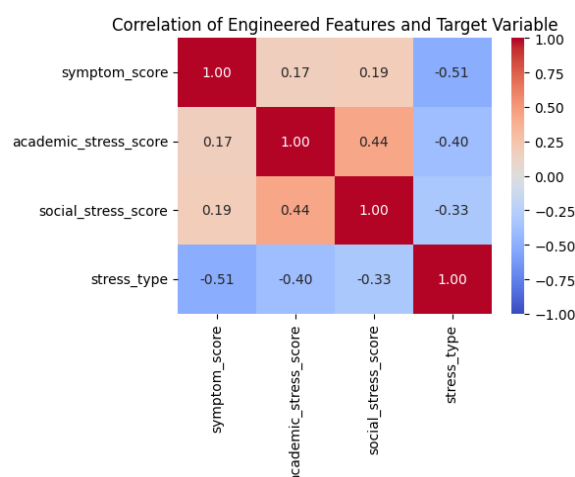


Figure 3. Correlation heatmap of engineered features and the target variable.

Figure 4 shows the distribution of stress categories after preprocessing. The Eustress category accounted for the majority of observations, whereas Distress and Other/Mixed Stress represented substantially smaller proportions of the dataset. This class imbalance is important when interpreting the classification results because a high accuracy value may partly reflect the dominance of the majority class. Therefore, balanced accuracy and macro-averaged metrics are also considered in the model evaluation.

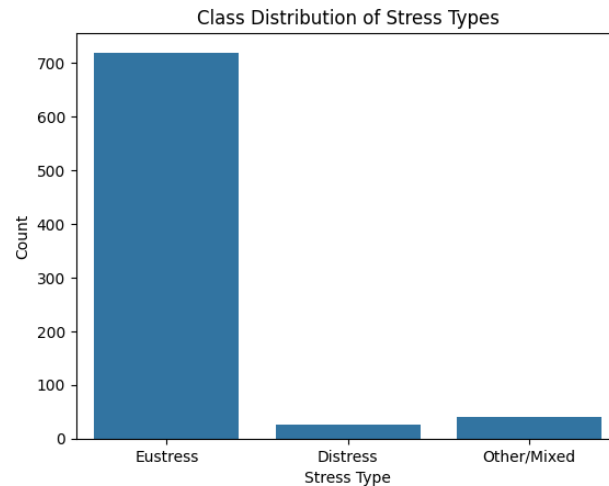


Figure 4. Distribution of stress type categories after preprocessing.

Additional insight is provided in Figure 5, which presents the correlation heatmap of the original predictor variables. Most feature pairs exhibited weak-to-moderate correlations, indicating limited multicollinearity among the predictors. Several stress-related variables, such as anxiety_tension, sadness_low_mood, concentration_problems, and loneliness_isolation, showed stronger relationships with stress_type than demographic variables. These findings suggest that psychological, academic, and social indicators contribute more substantially to stress classification than basic demographic characteristics alone. The results support the multidimensional nature of student stress, as the classification is associated with psychological symptoms as well as academic and social factors. Among the engineered features, symptom_score showed the strongest relationship with stress_type, followed by academic_stress_score and social_stress_score. This suggests that psychological symptoms may provide particularly useful information for distinguishing stress categories.

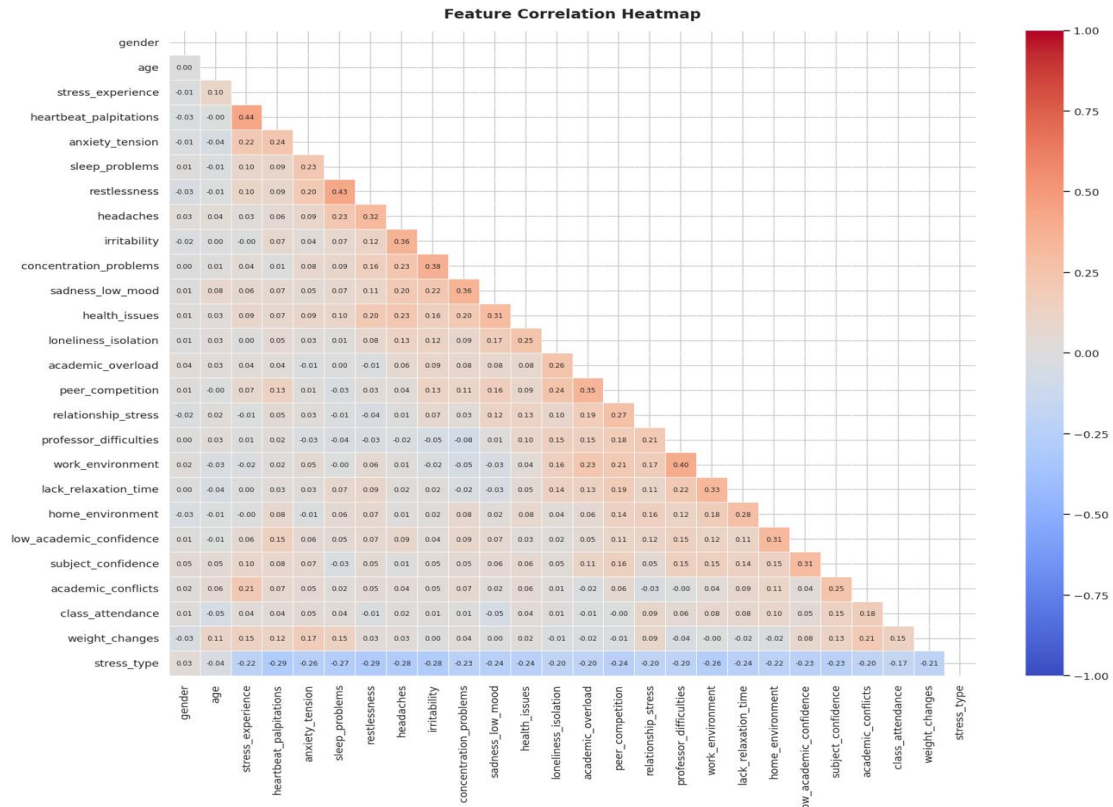


Figure 5. Correlation heatmap of original predictor variables.

3.2 Model Performance Results

Table 2 summarizes the performance of the three classification algorithms on the original and feature-engineered datasets. To provide a more reliable estimate of model performance, the results are reported as the mean and standard deviation

obtained from stratified five-fold cross-validation. The majority-class classifier is included as a baseline to determine whether the machine learning models provide meaningful improvement beyond the class distribution alone.

Table 2. Cross-Validation Performance of the Compared Models

Model	Accuracy	Balanced Accuracy	Macro Precision	Macro Recall	Macro F1	Weighted F1
Logistic Regression (Original)	0.9759 ± 0.0109	0.8199 ± 0.0765	0.9836 ± 0.0165	0.8199 ± 0.0765	0.8815 ± 0.0518	0.9734 ± 0.0125
Logistic Regression (Engineered)	0.9759 ± 0.0109	0.8199 ± 0.0765	0.9836 ± 0.0165	0.8199 ± 0.0765	0.8815 ± 0.0518	0.9734 ± 0.0125
KNN (Original)	0.9454 ± 0.0065	0.6105 ± 0.0472	0.9357 ± 0.0389	0.6105 ± 0.0472	0.7016 ± 0.0483	0.9354 ± 0.0087
KNN (Engineered)	0.9479 ± 0.0103	0.6161 ± 0.0755	0.9495 ± 0.0671	0.6161 ± 0.0755	0.7073 ± 0.0813	0.9380 ± 0.0141
Gaussian Naive Bayes (Original)	0.9555 ± 0.0040	0.8803 ± 0.0571	0.8033 ± 0.0231	0.8803 ± 0.0571	0.8342 ± 0.0290	0.9564 ± 0.0057
Gaussian Naive Bayes (Engineered)	0.9365 ± 0.0098	0.9145 ± 0.0533	0.7251 ± 0.0370	0.9145 ± 0.0533	0.7955 ± 0.0341	0.9422 ± 0.0081

LR achieved the highest accuracy among the evaluated machine learning models, with an accuracy of 97.59% ± 1.09% for both the original and engineered feature sets. It also achieved the highest macro F1-score of 88.15% ± 5.18%. Compared with the majority baseline of 91.36%, LR provides an improvement of approximately 6.23 percentage points in accuracy. This indicates that the high accuracy is not solely attributable to predicting the majority Eustress class.

KNN achieved an accuracy of 94.54% ± 0.65% using the original features and 94.79% ± 1.03% using the engineered features. The corresponding macro F1-scores were 70.16% and 70.73%, respectively. Although its overall accuracy was relatively high, its substantially lower macro F1-score indicates weaker performance across the minority classes. The small improvement after feature engineering suggests that the composite features provide limited additional benefit for the distance-based classifier.

GNB achieved 95.55% ± 0.40% accuracy with the original features and 93.65% ± 0.98% with the engineered features. Interestingly, GNB with engineered features achieved the highest balanced accuracy, at 91.45% ± 5.33%, indicating relatively strong recall across the three classes. However, its macro precision decreased to 72.51% ± 3.70%, resulting in a macro F1-score of 79.55% ± 3.41%. This indicates a trade-off between identifying minority classes and generating false-positive classifications.

The comparison also shows that feature engineering does not consistently improve all models. The engineered features produced a small improvement for KNN but did not change the LR results and reduced the overall accuracy of GNB. Therefore, the usefulness of feature engineering appears to depend on the learning mechanism of the classifier rather than providing a universal performance improvement.

A brief inspection of the classification errors shows that most misclassifications occurred between the minority classes and the majority Eustress class. Logistic Regression retained the highest overall performance, although some Distress and Other/Mixed cases were classified as Eustress. KNN showed a similar pattern with a higher proportion of minority-class misclassification, whereas Gaussian NB identified more minority-class cases but produced more false positives for the Eustress class. These results indicate that the high overall accuracy should be interpreted together with class-specific performance, particularly under the observed class imbalance.

Table 3. Per-Class Precision, Recall, and F1-Score of the Compared Models from Pooled Out-of-Fold Predictions

Model	Class	Precision	Recall	F1	Support
LR	Distress	1.000	0.704	0.826	27
LR	Eustress	0.976	0.999	0.987	719
LR	Other/Mixed	0.969	0.756	0.849	41
KNN	Distress	0.923	0.444	0.600	27
KNN	Eustress	0.947	0.996	0.971	719
KNN	Other/Mixed	0.889	0.390	0.542	41
GNB	Distress	0.727	0.889	0.800	27
GNB	Eustress	0.983	0.968	0.975	719
GNB	Other/Mixed	0.696	0.780	0.736	41

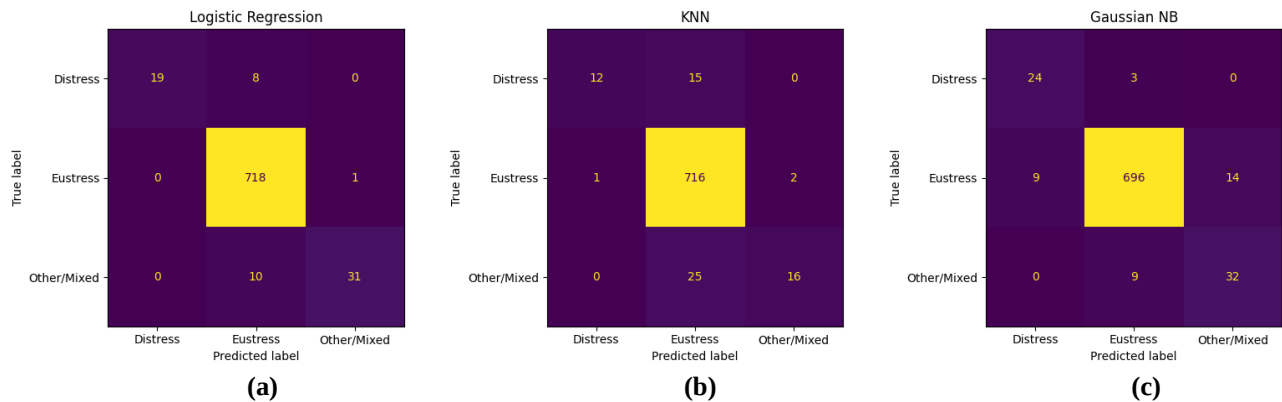


Figure 6. Confusion Matrices from Pooled Out-of-Fold Predictions for (a) Logistic Regression, (b) K-Nearest Neighbors, and (c) Gaussian Naive Bayes

Pooled out-of-fold predictions show that the dominant error for all models is a minority-class respondent being predicted as Eustress (Table 3 and Figure 6). LR misclassified 10 of 41 Other/Mixed and 8 of 27 Distress cases as Eustress, recalling 75.6% and 70.4% of these classes. KNN performed worst on the minority classes, with 25 of 41 Other/Mixed and 15 of 27 Distress cases predicted as Eustress (recall 39.0% and 44.4%). GNB recalled the most minority cases (88.9% Distress, 78.0% Other/Mixed) but produced 23 false alarms from Eustress respondents (14 as Other/Mixed, 9 as Distress), which lowered its precision to 72.7% and 69.6%. Because the minority classes contain only 27 and 41 samples, these class-level estimates are unstable, and a single case changes recall by 2–4 percentage points.

3.3 Discussion

The cross-validation results in Table 2 suggest that the three stress categories in this dataset can be classified effectively using a linear model, which helps explain why an unoptimized Logistic Regression (LR) achieved the highest overall performance, with $97.59\% \pm 1.09\%$ accuracy and $88.15\% \pm 5.18\%$ macro F1-score, compared with both a distance-based and a probabilistic classifier. However, this result should not be interpreted as evidence that the classes are universally or clinically linearly separable, because the observed performance is specific to the predictor structure and population represented in this dataset. This finding diverges from the direction taken by most recent literature on student stress classification, which tends to favor ensemble or tree-based methods over linear models. Omarbekova et al. compared five classification algorithms on academic stress data and reported accuracy values between 0.8727 and 0.8909, with Random Forest performing best overall, although LR still achieved the highest AUC-ROC score in their study, indicating that even in a dataset dominated by ensemble performance, the feature space retained meaningful linear separability [18]. Liu and Yu optimized a decision tree using the Harris Hawks Optimization algorithm and found that the base decision tree reached only 90.9% accuracy, improving to 92.7% after metaheuristic tuning, while Doma et al. reported an ensemble accuracy of 92.41% for a hybrid model combining five base classifiers on student stress-level data [19–20]. Compared with these results, the cross-validated 97.59% accuracy obtained here using an unoptimized LR remains comparatively high. Nevertheless, the majority-class baseline already achieved 91.36% accuracy, indicating that the class distribution contributes substantially to the high accuracy. The additional evaluation using balanced accuracy and macro F1 therefore provides a more informative basis for interpreting the model performance.

Not every recent study aligns with the strength of LR observed in this research. De Filippis and Foysal relied exclusively on Random Forest for a comparable stress-factor dataset of 1,100 students and reported a cross-validated accuracy of 0.88, while explicitly noting that a single-algorithm, feature-importance-focused approach limits the explainability of how specific factors interact with one another [21]. Reading alongside the present findings, this contrast suggests that the contribution of LR in this study lies not only in its accuracy but in the direct interpretability of its coefficients, since comparable or higher accuracy can be reached without ensemble complexity when the original feature set already carries strong discriminative information. The present cross-validation results further support this interpretation, as LR maintained the same performance for both the original and engineered feature sets, suggesting that the original predictors already contain substantial information for the classification task.

The divergent effect of feature engineering across the three algorithms is also consistent with observations elsewhere in educational data mining. Chen et al. noted that Naive Bayes models tend to degrade when the assumed conditional independence between features does not hold, since real-world educational and behavioral features are typically correlated [22]. This supports the explanation proposed for the decline in GNB accuracy after the composite features were introduced: because symptom_score, academic_stress_score, and social_stress_score were constructed from variables already present in the original feature set, they likely introduced redundancy that conflicts with the independence assumption underlying GNB, whereas the same redundancy appears to have provided a more useful notion of similarity for the distance-based K-Nearest Neighbors classifier. The cross-validation results provide partial support for this interpretation: KNN showed a small improvement in accuracy from 94.54% to 94.79% and in macro F1 from 70.16% to 70.73% after feature engineering, whereas GNB accuracy decreased from 95.55% to 93.65%. However, the engineered GNB model achieved the highest balanced accuracy ($91.45\% \pm 5.33\%$), indicating that the reduction in overall

accuracy was accompanied by improved average recall across classes. Most of the studies cited above focus on identifying a single best-performing algorithm on a fixed feature representation rather than examining how one feature engineering strategy affects multiple algorithms differently; the present study addresses this gap directly by evaluating the same engineered features across three classifiers with distinct inductive biases, offering a more algorithm-aware perspective than the existing literature.

The class imbalance visible in Figure 4, where Eustress substantially outnumbers Distress and Other/Mixed Stress, has direct implications for how these results should be read. The majority-class baseline of 91.36% demonstrates that accuracy alone would provide an overly optimistic assessment of classification performance. Therefore, the present study additionally reports balanced accuracy and macro-averaged metrics to assess performance across the three classes. Because weighted precision, recall, and F1-score were used rather than accuracy alone, the reported performance is less likely to reflect majority-class dominance; however, weighted metrics still give greater influence to the majority Eustress class and therefore cannot fully represent minority-class recognition. The cross-validation results illustrate this limitation: KNN achieved 94.79% accuracy but only 70.73% macro F1-score, whereas LR achieved 97.59% accuracy and a substantially higher macro F1-score of 88.15%. The model's ability to generalize to students whose stress profile differs from the dominant Eustress pattern remains uncertain. A brief inspection of classification errors also showed that minority Distress and Other/Mixed cases were more frequently misclassified as Eustress, while GNB showed a different trade-off by identifying more minority cases at the cost of additional false-positive predictions. This limitation is particularly relevant given the practical objective of this study, since a screening tool intended to flag students who may need intervention is arguably most useful precisely for the minority Distress and Other/Mixed Stress cases that require careful evaluation beyond overall accuracy.

Taken together, these findings extend the research direction identified in Section 1 by showing that a simple, interpretable linear model can achieve strong overall performance relative to the more complex approaches reported in recent studies [18-20], provided the underlying feature set already carries strong discriminative information, and that the value of feature engineering is contingent on how well the transformed representation aligns with each algorithm's assumptions rather than being universally beneficial. However, direct comparison with these studies is limited because they used different datasets and labels. The comparison also indicates that no single model dominates all evaluation criteria: LR provides the strongest overall accuracy and macro F1-score, whereas GNB with engineered features provides the highest balanced accuracy. Thus, model selection should be guided by the intended objective rather than accuracy alone. For institutions with limited technical infrastructure, this suggests that investment in well-chosen, psychologically grounded predictor variables may offer a better return than investment in model complexity, although the findings should be considered dataset-specific and should not be interpreted as evidence of clinical validity or broad generalizability. Any deployment should treat the classifier's output as a screening aid rather than a diagnostic judgment, consistent with the broader caution raised in prior systematic reviews of machine learning in student mental health [9-10]. Future work could extend this study by applying class-balancing techniques such as the SMOTE to address the observed imbalance between Eustress, Distress, and Other/Mixed Stress, by comparing these interpretable models against deep learning architectures, which have recently shown strong performance in screening anxiety, depression, and stress among university students, and by exploring alternative composite indicators similar to holistic psychological scores used in related student mental health classification work [23-25].

4. CONCLUSION

This study evaluated LR, KNN, and GNB for classifying student stress types. Based on five-fold cross-validation, LR achieved the highest overall performance, with $97.59\% \pm 1.09\%$ accuracy and $88.15\% \pm 5.18\%$ macro F1-score using both the original and engineered feature sets. However, GNB with engineered features achieved the highest balanced accuracy of $91.45\% \pm 5.33\%$, indicating that model performance varies depending on the evaluation criterion and the importance given to minority-class recognition. Compared with the majority-class baseline of 91.36% accuracy, the results indicate that the machine learning models provide additional predictive value beyond the dominant Eustress class, although the class imbalance remains an important consideration. The results do not imply that the three stress categories are universally or clinically distinct, but rather indicate that they can be effectively classified within the evaluated dataset using relatively simple machine learning approaches. The feature engineering process, which produced the symptom score, academic stress score, and social stress score composite variables, gave a modest performance improvement for the distance-based KNN model but did not enhance, and in the case of GNB actually reduced, predictive performance, indicating that the value of a given feature representation depends heavily on the assumptions underlying each algorithm rather than being universally beneficial. The composite scores nevertheless provide a more concise representation of related questionnaire items and may support interpretation of broader symptom, academic, and social dimensions of student stress. Although the dataset exhibited a substantial class imbalance dominated by the Eustress category, the use of balanced accuracy and macro-averaged metrics provided a more informative assessment than accuracy alone. The error analysis further showed that minority Distress and Other/Mixed Stress cases were more frequently classified as Eustress, highlighting the importance of considering class-specific performance when interpreting the results. Accordingly, LR can be considered a computationally lightweight and interpretable model for this dataset, but its performance should be viewed as a screening-oriented result rather than evidence of clinical applicability or diagnostic capability. Several limitations

should be acknowledged, including class imbalance and the use of a single public dataset, which may limit the generalizability of the findings across student populations and institutional contexts. The high predictive performance may also reflect dataset-specific characteristics, and external validation on independent and more diverse student populations is necessary before broader application can be considered. Future research could examine class-balancing techniques such as oversampling, compare deep learning architectures with interpretable baseline models, and validate the proposed approach using larger and more demographically diverse student populations.

5. DECLARATIONS

5.1 Author Contributions

Conceptualization: author1, author2 and author3;
Methodology: author1, author2 and author3;
Software: author1 and author2;
Validation: author3;
Formal Analysis: author1, author2 and author3;
Investigation: author1 and author2;
Resources: author3;
Data Curation: author1 and author2;
Writing Original Draft Preparation: author1 and author2;
Writing Review and Editing: author1, author2 and author3;
Visualization: author1 and author2;
All authors have read and agreed to the published version of the manuscript.

5.2 Data Availability

The data presented in this study are available on request from the corresponding author.

5.3 Funding

The publication and manuscript processing fees associated with the preparation and publication of this scientific manuscript are supported by the Internal Student Research Fund Program provided by Lembaga Penelitian dan Pengabdian Masyarakat (LPPM), Universitas Ciputra Surabaya.

5.4 Institutional Review Board Statement

Not applicable.

5.5 Informed Consent Statement

Not applicable.

5.6 Declaration of Competing Interest

The authors declare that they have no competing financial interests or personal relationships.

5.7 Generative AI and AI-assisted Technologies in The Writing Process

The authors used Generative AI to improve the clarity and readability of this paper. All AI-assisted content was reviewed and edited by the authors, who take full responsibility for the final manuscript.

REFERENCES

- [1] Gallup, "One in Three College Students Consider Leaving Program," Gallup News, Apr. 30, 2025. [Online]. Available: <https://news.gallup.com/poll/659897/one-three-college-students-consider-leaving-program.aspx>
- [2] E. Karyotaki et al., "Sources of Stress and Their Associations With Mental Disorders Among College Students: Results of the World Health Organization World Mental Health Surveys International College Student Initiative," *Frontiers in Psychology*, vol. 11, Art. no. 1759, 2020, doi: 10.3389/fpsyg.2020.01759.
- [3] M. Pascoe, S. Hetrick, and A. Parker, "The impact of stress on students in secondary school and higher education," *International Journal of Adolescence and Youth*, vol. 25, pp. 104–112, 2020, doi: 10.1080/02673843.2019.1596823.
- [4] B. Hives, M. R. Beauchamp, Y. Liu, J. Weiss, and E. Puterman, "Multidimensional correlates of psychological stress: Insights from traditional statistical approaches and machine learning using a nationally representative Canadian sample," *PLOS ONE*, vol. 20, Art. no. e0323197, 2025, doi: 10.1371/journal.pone.0323197.
- [5] H. Selye, *The Stress of Life*, rev. ed. New York, NY, USA: McGraw-Hill, 1976.
- [6] R. Upsher et al., "Aligning interventions with the University Mental Health Charter: A stratified review of reviews of mental health and wellbeing interventions for higher education students," *Educational Research Review*, vol. 49, Art. no. 100716, 2025, doi: 10.1016/j.edurev.2025.100716.
- [7] A. B. R. Shatte, D. M. Hutchinson, and S. J. Teague, "Machine learning in mental health: A scoping review of methods and applications," *Psychological Medicine*, vol. 49, no. 9, pp. 1426–1448, 2019, doi: 10.1017/S0033291719000151.



- [8] R. V. Anand, A. Q. Md, S. Urooj, S. Mohan, M. A. Alawad, and A. C., "Enhancing Diagnostic Decision-Making: Ensemble Learning Techniques for Reliable Stress Level Classification," *Diagnostics*, vol. 13, no. 22, Art. no. 3455, 2023, doi: 10.3390/diagnostics13223455.
- [9] B. L. Schaab, P. Ü. Calvetti, C. T. Reppold, et al., "How do machine learning models perform in the detection of depression, anxiety, and stress among undergraduate students? A systematic review," *Cadernos de Saúde Pública*, vol. 40, no. 12, 2024, doi: 10.1590/0102-311XEN029323.
- [10] A. Daza, N. Saboya, J. I. Necochea-Chamorro, K. Z. Ramos, and Y. R. V. Valencia, "Systematic review of machine learning techniques to predict anxiety and stress in college students," *Informatics in Medicine Unlocked*, vol. 43, Art. no. 101391, 2023, doi: 10.1016/j.imu.2023.101391.
- [11] C. Ruiz-Camacho and M. Gozalo, "Predicting University Students' Stress Responses: The Role of Academic Stressors and Sociodemographic Variables," *European Journal of Investigation in Health, Psychology and Education*, vol. 15, no. 8, Art. no. 163, 2025, doi: 10.3390/ejihpe15080163.
- [12] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.
- [13] R. Chang, N. Kuswanto, J. Tedja, and C. Andreas, "Penerapan regresi logistik, K-NN, dan Naïve Bayes berbasis pendekatan CRISP-DM dalam memprediksi penyakit jantung," *Building of Informatics, Technology and Science (BITS)*, vol. 7, no. 4, pp. 2771–2779, 2026, doi: 10.47065/bits.v7i4.8518.
- [14] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. New York, NY, USA: Springer, 2009, doi: 10.1007/978-0-387-84858-7.
- [15] A. H. Alshammari, G. Bencsik, and H. A. Almashhadani, "A Survey of Six Classical Classifiers, Including Algorithms, Methodological Characteristics, Foundational Variants, and Recent Advances," *Algorithms*, vol. 19, no. 1, Art. no. 37, 2026, doi: 10.3390/a19010037.
- [16] M. M. R. Mondol and M. A. Kabir, *Stress indicators dataset for mental health classification*, ver. 2, Mendeley Data, 2025, doi: 10.17632/2gsjv8m7ch.2.
- [17] F. Sword and C. Andreas, "Credit card fraud detection under extreme class imbalance: A comparison of KNN and logistic regression," *UNP Journal of Statistics and Data Science*, vol. 4, no. 3, pp. 417–424, 2026, doi: 10.24036/ujsds/vol4-iss3/539.
- [18] A. Omarbekova, A. Nazyrova, G. Bekmanova, Z. Lamasheva, A. Gibadullina, and A. Abdikarim, "An explainable machine learning framework for academic stress classification among university students: a comparative multi-dataset study with ablation analysis and statistical significance testing," *Front. Comput. Sci.*, vol. 8, 2026, Art. no. 1886274, doi: 10.3389/fcomp.2026.1886274.
- [19] C. Liu and S. Yu, "Students' stress prediction and explainable analysis based on improved decision trees," *Front. Psychol.*, vol. 16, 2026, Art. no. 1684529, doi: 10.3389/fpsyg.2025.1684529.
- [20] V. Doma, A. Abd Almisreb, E. Yaman, S. Amanzholova, and N. Ismail, "Classification of student stress levels using a hybrid machine learning model," in *AI-Driven: Social Media Analytics and Cybersecurity*, Studies in Computational Intelligence, vol. 1180, W. M. Yafooz and Y. Al-Gumaei, Eds. Cham: Springer, 2025, pp. 77–104, doi: 10.1007/978-3-031-80334-5_5.
- [21] R. de Filippis and A. A. Foysal, "Comprehensive analysis of stress factors affecting students: a machine learning approach," *Discov. Artif. Intell.*, vol. 4, no. 1, p. 62, 2024, doi: 10.1007/s44163-024-00169-6.
- [22] Y. Chen, J. Sun, J. Wang, L. Zhao, X. Song, and L. Zhai, "Machine learning-driven student performance prediction for enhancing tiered instruction," *arXiv:2502.03143*, 2025.
- [23] D. Elreedy, A. F. Atiya, and F. Kamalov, "A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning," *Mach. Learn.*, vol. 113, no. 7, pp. 4903–4923, Jul. 2024, doi: 10.1007/s10994-022-06296-4.
- [24] B. Juárez-Santiago, K. Olvera-Raymundo, J. M. Olivares-Ramírez, N. Olguín-López, O. Rodríguez Abreo, and J. Rodríguez-Reséndiz, "Integrating Deep Learning into Educational Wellbeing: Early Screening of Anxiety, Depression, and Stress Among University Students," *Educ. Sci.*, vol. 16, no. 1, Art. no. 50, 2026, doi: 10.3390/educsci16010050.
- [25] R. H. Tanjung, F. Damayanti, and A. Zaki, "Optimasi Random Forest Melalui Feature Engineering dan SMOTE untuk Klasifikasi Kesehatan Mental," *Bulletin of Information Technology (BIT)*, vol. 7, no. 2, pp. 204–212, 2026, doi: 10.47065/bit.v7i2.2722.