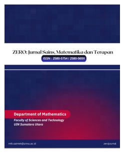
Zero : Jurnal Sains, Matematika, dan Terapan
E-ISSN : 2580-5754; P-ISSN : 2580-569X
Volume 9, Number 1, 2025
DOI: 10.30829/zero.v9i1.24537
Page: 264-276



Predicting Student Stress Levels Based on Lifestyle Factors Using the Catbost Algorithm

[1] **Putri Meuthia Rani**
Department of Computer Science, UIN Sumatera Utara Medan, Medan Indonesia

[2] **Ilka Zufria**
Department of Computer Science, UIN Sumatera Utara Medan, Medan Indonesia

Article Info

Article history:

Accepted, 28 June 2025

Keywords:

CatBoost Algorithm;
Lifestyle Factors;
Mental Health;
Stress Prediction;
Student Stress

ABSTRACT

This study developed a machine learning model to classify student stress levels based on lifestyle factors using the CatBoost algorithm. Data were collected from 630 students of the SciTech Faculty at State Islamic University of North Sumatra through a questionnaire comprising 14 Likert-scale items. Instrument validation was confirmed using Pearson's r (>0.821, $p < 0.05$) and Cronbach's Alpha (0.866). Preprocessing included outlier removal with IQR, feature encoding, stratified train-test split (80:20), and 5-fold cross-validation. The training set was imbalanced and addressed using the SMOTE technique. Model evaluation used accuracy (85%), precision, recall, and F1-score per class, with high recall (0.97) for moderate and improved F1-score (0.79) for low stress. Final classification used a 20% test subset (126 samples). Feature importance analysis identified task procrastination, poor sleep quality, and weak time management as key predictors. These findings affirm CatBoost's reliability through consistent results, scalability, and balanced evaluation metrics beyond mere accuracy.

Corresponding Author:

Putri Meuthia Rani,
Department of Computer Science,
State Islamic University of North Sumatra, Medan, Indonesia
Email: meuthiaray@gmail.com

1. INTRODUCTION

Mental health, as described by the World Health Organization (WHO), is a state that supports optimal individual development physically, mentally, and emotionally, as well as harmony with the surrounding environment; not only the absence of disorders, but also the ability to manage stress, work productively, and contribute to society [1]–[5]. Mental health is an important global issue, particularly among the younger generation. Students, who belong to this demographic group, experience a critical transitional phase that involves complex academic and social expectations. Several studies have shown that university students face psychological stress due to workload, performance pressure, and organizational responsibilities [6]–[8].

For instance, report that academic burden and extracurricular roles significantly contribute to mental strain among undergraduate populations in Indonesia. These demands, when exceeding one's adaptive threshold, lead to stress responses characterized by cognitive, emotional, and behaviour disruptions [9], [10]. Further studies emphasize that unmanaged stress among students can result in anxiety, burnout, and reduced academic performance. This study focuses on these challenges by predicting student stress levels based on lifestyle-related factors, as they have been shown to be modifiable indicators that influence one's mental state under academic strain [11]–[13].

In the Faculty of Science and Technology, students often face stress due to high academic demands, such as practicum, practicum reports, and various course assignments with tight collection schedules that require completion in a limited time while understanding the material in depth. This increasing academic load often causes students to feel overwhelmed, plus lifestyle factors such as irregular sleep patterns, lack of physical activity, and consumption of unhealthy foods can worsen stress conditions [14]–[18]. Data from the Ministry of Health of the Republic of Indonesia shows that approximately 30% of university students in Indonesia have experienced high levels of stress due to academic pressure, underscoring the need to identify triggering factors and appropriate coping strategies[19].

This research utilizes the CatBoost algorithm due to its advantage in efficiently processing categorical data without extensive preprocessing, such as one-time coding. CatBoost is also known for its ability to reduce the risk of overfitting and produce high accuracy even when working with relatively small data sets[20]–[25]. The optimized gradient boosting framework enables CatBoost to effectively handle datasets with complex structures and strong inter-variable correlations, making it particularly suitable for uncovering hidden relationships between lifestyle factors and student stress levels [26]–[30].

Stress in university students has a significant impact on mental health and academic performance. Lifestyle factors such as time management, sleep patterns and physical activity are known to influence stress levels. Along with technological developments, *machine learning* approaches have begun to be utilized to predict stress levels based on various indicators. The K-Nearest Neighbour (KNN) algorithm was used to predict UNUGHA student stress based on study habits and psychological conditions, achieving 83.33% accuracy, although the study utilized limited data and did not test more complex algorithms [31]. Another study compared several classification algorithms on student data, finding Perceptron to be the most accurate, but it did not explore advanced strategies[32]. Stress classification based on physiological data during sleep was developed using Random Forest, but lifestyle factors were not thoroughly considered [33].

Although numerous studies have investigated stress prediction among university students, only a few have specifically explored the relationship between lifestyle factors and the CatBoost algorithm. In fact, CatBoost is particularly effective in processing categorical data, such as lifestyle survey responses. Therefore, this study aims to fill this gap. The research question is: *Can the CatBoost algorithm classify university students' stress levels based on lifestyle factors?* Accordingly, the objective of this study is to develop a lifestyle-based stress prediction model for university students using the CatBoost algorithm with high accuracy.

2. RESEARCH METHOD

This study adopts a quantitative approach to analyze the relationship between lifestyle factors and student stress levels at the Faculty of Science and Technology, State Islamic University of North Sumatra, for the 2021–2023 cohorts. A total of 630 student data points were collected using purposive sampling to ensure that the participants were relevant to the research objectives.

The research stages include problem identification, data collection, data preprocessing, model development using the CatBoost algorithm, and model evaluation. All analyses were conducted using Google Colab as the programming environment. The CatBoostClassifier was selected due to its ability to efficiently process categorical features without requiring one-hot encoding, and its strong generalization performance on small datasets. Compared to other ensemble methods such as XGBoost and Random Forest, CatBoost demonstrates better resistance to overfitting through its ordered boosting approach. Several key hyperparameters were tuned to enhance model performance. The following hyperparameters were used:

1. Iterations = 200 (to avoid overfitting).
2. Learning rate = 0.1 (for balanced convergence and generalization).
3. Depth = 6 (to control model complexity).
4. Random_seed = 42 (for reproducibility).

Hyperparameter tuning was conducted manually through iterative testing, with accuracy and F1-score on validation data as performance metrics. This approach provides flexibility in selecting the optimal configuration for the dataset.

2.1 Data Collection

Primary data were obtained from 630 students of the Faculty of Science and Technology through a questionnaire. The questionnaire was developed based on preliminary interviews with psychology experts and student representatives to ensure the relevance and coverage of the variables. The questionnaire employed a 4-point Likert scale (1 = Never, 2 = Rarely, 3 = Sometimes, 4 = Often) to assess various lifestyle factors and indicators of stress levels. Prior to distribution, the instrument was tested for validity and reliability.

Table 1. Response Category Scale and Descriptions

Score	Category
1	Never

2	Rarely
3	Sometimes
4	Often

The Pearson correlation validity test showed all 14 items had r-values > 0.821 and p-values < 0.05, indicating strong and significant item validity. If the values were below these thresholds, the items would require revision. Reliability was excellent with a Cronbach's Alpha of 0.866.

2.2 Data Pre-processing

The raw data obtained from the questionnaire then went through a series of pre-processing stages to ensure the quality and readiness of the data before modelling. These stages include:

1. Data Cleaning: Handling of missing values is done to maintain the integrity of the *dataset.*
2. Data Normalization: Data normalization is an essential process in machine learning to ensure that the scale of feature values is uniform. In this study, normalization was performed before inputting the data into the CatBoost algorithm to prevent the model from being biased toward features with larger scales. Figure 2 illustrates the steps involved in the data normalization process used in this study.

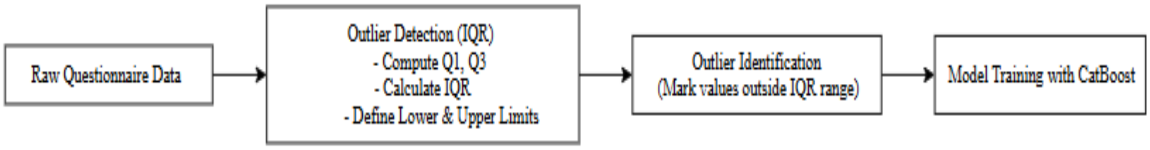


Figure 2. Flowchart of the data normalization process before applying the CatBoost model

3. Outliers were detected using the Interquartile Range (IQR) method on all attributes (Q1–Q14) to prevent distortion in the analysis. Although all values fell within the 1–4 Likert scale, some outliers were identified in Q5, Q7, Q9, Q10, and Q14. These were removed to maintain data quality. The results of IQR-based outlier detection are presented below:

Table 2. IQR calculation

Attributes	Q1	Q3	IQR	Lower Limit	Upper Limit	Outlier Detected
Q1	2.0	3.0	1.0	0.5	4.5	0
Q2	2.0	3.0	1.0	0.5	4.5	0
Q3	2.0	4.0	2.0	-1.0	7.0	0
Q4	1.0	2.0	1.0	-0.5	3.5	0
Q5	3.0	4.0	1.0	1.5	5.5	1
Q6	1.0	4.0	3.0	-3.5	8.5	0
Q7	3.0	4.0	1.0	1.5	5.5	1
Q8	2.0	3.0	1.0	0.5	4.5	0
Q9	1.0	2.0	1.0	-0.5	3.5	1
Q10	1.0	2.0	1.0	-0.5	3.5	1
Q11	2.0	4.0	2.0	-1.0	7.0	0
Q12	1.0	3.0	2.0	-2.0	6.0	0
Q13	1.0	3.0	2.0	-2.0	6.0	0
Q14	1.0	2.0	1.0	-0.5	3.5	1

4. Variable Encoding: Categorical variables were converted into numeric format, even though CatBoost can handle them natively.
5. Data Splitting and Validation: The dataset was divided using the train-test-split technique with a ratio of 80:20, where 80% of the data was used for model training and the remaining 20% for testing. To enhance reliability and reduce the risk of overfitting, additional validation was conducted using the 5-fold cross-validation technique. The parameter random_state=42 was employed to ensure the reproducibility of the results.

2.3 Modelling with CatBoost Algorithm

The prediction of student stress levels in this study was carried out using the CatBoost algorithm, an ensemble method based on Gradient Boosting Decision Trees (GBDT). CatBoost was chosen due to its strong capability in efficiently handling categorical features without requiring complex preprocessing such as one-hot encoding. Most of the lifestyle indicators in this study—such as sleep schedules, eating habits, physical activity, and social media usage—are categorical data, making CatBoost highly suitable. Furthermore, CatBoost is designed to perform optimally on medium-sized tabular datasets like the one used in this study (630 entries), with built-in features for handling missing values and minimizing the risk of overfitting through its ordered boosting approach. The implementation is carried out using the Python programming language and the CatBoost library. Before the model training process, the dataset underwent a data normalization procedure to ensure consistent feature scaling and improve model performance. The normalization process is illustrated in Figure 2.

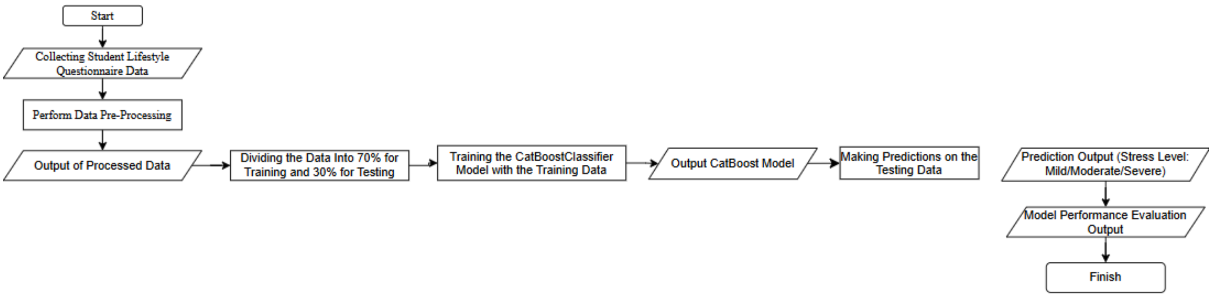


Figure 2. CatBoost-Based Stress Prediction Flowchart

The pre-processed dataset is divided into training and testing sets with a common ratio of 80:20. The model is trained on the training data and its performance is evaluated using the testing data. The evaluation metric used was MultiClass loss, with the best result obtained at the 1289th iteration. These results indicate that the selected hyperparameter configuration achieved optimal performance for the multiclas s classification of student stress levels.

2.4 Model Evaluation

Model performance was evaluated using accuracy and F1-score to assess overall classification effectiveness. A confusion matrix was used to visualize the distribution of correct and incorrect predictions across stress levels (mild, moderate, and severe). To ensure balanced class representation during data splitting, a stratified train-test split with an 80:20 ratio was applied, allocating 126 out of 630 entries for testing. This proportion is deemed representative of the original class distribution, allowing for a more objective and fair assessment. Based on the *confusion matrix*, several key evaluation metrics were calculated to assess the performance of the model:

1. **Accuracy** represents the overall correctness of the model's predictions by comparing the number of correct classifications to the total number of data instances [13], [25]. It is defined as:

$$\text{Accuracy}=\frac{\text{Number of correct predictions}}{\text{Total number of data instances}} \tag{1}$$

2. **Precision** measures the proportion of true positive predictions (students correctly identified as high stress) to all instances predicted as high stress. It reflects the model's ability to minimize false positives [1], [26]:

$$\text{Precision}=\frac{TP}{TP+FP} \tag{2}$$

3. **Recall (Sensitivity)** quantifies the proportion of actual high-stress students that were correctly identified by the model, reflecting its capacity to avoid false negatives [25], [34]:

$$\text{Recall}=\frac{TP}{TP+FN} \tag{3}$$

4. **F1-score** is the harmonic mean of precision and recall, offering a balanced assessment when there is a trade-off between the two. It is especially useful in imbalanced datasets [13], [23]:

$$\text{F1}=2 \text{ x } \frac{\text{Precision x Recall}}{\text{Precision+Recall}} \tag{4}$$

Where:
a. TP (True Positive) : High-stress students correctly predicted as high stress.
b. TN (True Negative): Non-high-stress students correctly predicted as not high stress.
c. FP (False Positive): Non-high-stress students incorrectly predicted as high stress.
d. FN (False Negative): High-stress students incorrectly predicted as not high stress.

3. RESULT AND ANALYSIS

3.1 Data Collection and Respondent Characteristics

Primary data were collected through the distribution of questionnaires to students of the Faculty of Science and Technology, State Islamic University of North Sumatra, during the period 2021–2023. A total of 630 respondents with diverse academic backgrounds and lifestyles participated. The dataset includes demographic attributes (gender, major, and class), 14 lifestyle-related items, and stress level labels determined based on the questionnaire's rating scale.

Table 3. Questionnaire Questions Related to Student Lifestyle

No.	Code	Question
1	Q1	I often sleep late even if there is no urgent need.

2	Q2	I feel like my sleep is not enough to keep me refreshed when I wake up.
3	Q3	I eat fast food or instant food more often than healthy food.
4	Q4	I haven't done much exercise or physical activity in the past week.
5	Q5	I often put off doing my assignments until close to the deadline.
6	Q6	I feel that I rarely interact or meet face-to-face with friends or family.
7	Q7	I feel that the load of coursework in one week is too heavy to complete.
8	Q8	I often feel overwhelmed by my lack of ability to organize my daily schedule.
9	Q9	I often do several tasks at once and feel unfocused.
10	Q10	I feel academically stressed because of the many assignments and exams in my major.
11	Q11	I feel that there is a very high level of competition among the students in my major.
12	Q12	I feel like I don't have enough time for non-academic activities because I focus on my major.
13	Q13	I felt pressured to prepare myself for my career after graduating from this major.
14	Q14	I feel that it takes a lot of time to learn and understand the material.

Table 4 provides a detailed description of the respondents' characteristics, including gender, class level, and major, to give an overview of the demographic distribution in this study.

Table 4. Number of Respondents Based on Gender

No.	Gender	Total	Percentage
1	Female	353	56,0%
2	Male	277	44,0%
3	Total	630	100%

Based on Table 4, the majority of respondents were female (56.0%), while 44.0% of respondents were male.

Table 5. Number of Respondents by Generation

No.	Force	Total	Percentage
1	2021	221	35,1%
2	2022	193	30,6%
3	2023	216	34,2%
4	Total	630	100%

Table 5 shows the distribution of respondents by batch, indicating relatively balanced participation from the 2021 to 2023 cohorts, with batch 2021 having the highest proportion of respondents (35.1%).

Table 6. Number of Respondents Based on Major

No.	Major	Total	Percentage
1	Computer Science	171	27,1%
2	Information System	124	19,7%
3	Biology	119	18,9%
4	Math	108	17,1%
5	Physics	108	17,1%
6	Total	630	100%

Although the data were collected from various study programs within the Faculty of Science and Technology, respondents were dominated by students from the Computer Science program (27.1%). Other programs, such as Information Systems, Biology, Mathematics, and Physics, showed relatively balanced proportions. It is important to note that the findings are limited to the context of students within this faculty and may not represent those from other faculties at the State Islamic University of North Sumatra. This limitation should be considered when interpreting the results and designing future studies.

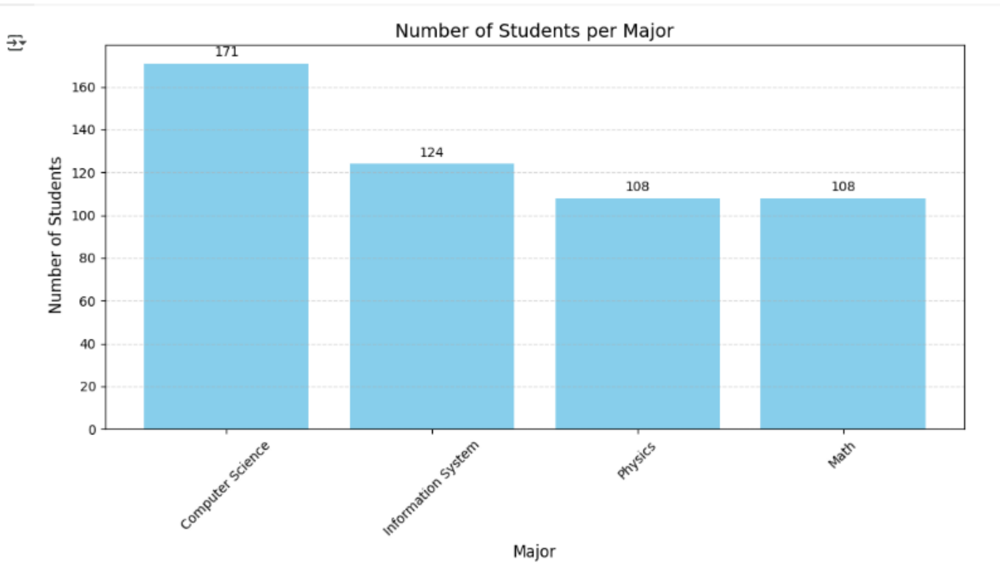


Figure 3. Student Major Distribution

3.2 Data Pre-processing Results

The data pre-processing stage ensures that the dataset is clean, consistent, and ready for model training. The key steps include:

1. Uploading and reading the excel-based questionnaire data using the pandas library in Google Colab.
2. Selecting 14 lifestyle-related columns (Q1–Q14) as input features.
3. Creating a 'Total_Score' column by summing all statements, then labelling stress levels as Low (below 10th percentile), High (above 90th), and Medium (in between). This percentile-based labelling approach enables adaptive thresholds based on the actual score distribution, commonly used when standardized cutoffs are not available. It ensures that the classification reflects relative stress intensity within the dataset [1].
4. Converting categorical attributes into numerical form, such as Gender: 0 = Female, 1 = Male; Major: 1 = Physics, 2 = Computer Science, 3 = Mathematics, 4 = Information Systems, 5 = Biology.
5. Checking for missing values; the dataset was complete, so no imputation was needed.

Table 7. Example of Data Pre-processing Results

JK	THE COURT	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14
0	2	2	3	2	2	2	1	4	2	3	2	2	1	3	2
0	2	4	3	4	4	1	1	2	1	2	1	2	2	3	4
0	2	2	2	2	2	2	2	3	2	2	2	2	2	2	2
0	2	2	2	3	4	4	3	4	4	4	4	3	3	4	4
0	2	3	3	3	3	3	2	3	3	3	4	4	4	4	4
0	1	1	2	2	2	2	1	4	3	3	3	2	3	2	2
0	3	2	2	1	1	1	1	1	1	2	2	2	1	1	4
0	2	3	4	3	2	1	1	2	1	1	1	2	1	3	2
0	2	3	3	2	4	3	1	2	2	2	2	3	4	4	4
...	...	..	...	...	...	...	...	...	...	...	...	...	...	...	...
1	4	1	3	2	1	2	2	2	2	3	3	4	2	4	3

After labelling, the dataset was divided into training and testing subsets using a stratified split with an 80:20 ratio. This method preserved the proportion of stress level classes in both subsets. The training set consisted of 504 samples, while the testing set included 126 samples.

Initial analysis revealed that the training data was highly imbalanced, with the following class distribution: Medium = 317 instances; Low = 167 instances; High = 20 instances.

To address this, the Synthetic Minority Over-sampling Technique (SMOTE) was applied exclusively to the training data. SMOTE generates synthetic samples for the minority classes by interpolating between existing instances, helping to reduce model bias, or this balancing step ensured that the classifier could learn equally from all classes, particularly the underrepresented High stress category. After applying SMOTE, the training class distribution became balanced as follows: Medium = 317 instances; Low = 317 instances; High = 317 instances.

3.3 Modelling Results with CatBoost Algorithm

The CatBoost algorithm was used to train the model on 504 training samples. Parameters such as random_seed ensured result reproducibility, and verbose=0 suppressed detailed output. The training aimed to build a model that could generalize patterns from training to test data. A manual calculation using 20 training samples illustrates how CatBoost works (target y: 0 = Low, 1 = Medium, 2 = High).

Table 7. Sample Training Data (20 Data) for Manual Calculation

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	y
1	4	3	3	3	4	3	3	1	1	4	1	2	1	1
2	2	1	1	1	1	1	1	2	2	2	1	1	4	0
3	3	4	2	4	4	4	3	2	4	1	3	4	2	2
4	3	2	2	4	2	3	4	4	1	4	4	1	1	1
5	2	4	1	3	1	4	2	1	1	2	2	3	1	1

1. Initial probability is computed using softmax: $P_{ik} = \frac{e^0}{e^0 + e^0 + e^0} = \frac{1}{3} = 0.333$
2. Calculation of iteration 1
   a. Calculate the residual (gradient): $g_{ik} = p_{ik} - 1_{[y1=k]} = \frac{1}{3} - 1_{[y1=k]}$

No.	y	g_{ik}
1	1	0.333 - 1 = -0.667
2	0	0.333 - 0 = 0.333
3	2	0.333 - 2 = 1.667
4	1	-0.667
5	0	0.333-0=0.333

   b. Score update iteration 1 (Scores are updated using the learning rate and gradient): $S^{(1)}{}_{ik} = S^{(0)}{}_{ik} - \eta\, F^{(0)}{}_{ik} = 0 - \eta . g^{(0)}{}_{ik}$ with learning rate = 0.1 to.$S^{(1)}{}_{ik} = -0.1 \text{ x } g^{(0)}{}_{ik}$

No.	y	g_{ik}	$S^{(1)}{}_{ik}$
1	1	-0.667	-0.1 x (-0.667) = 0.0667
2	0	0.333	-0.1 x 0.333 = -0.0333
3	2	1.667	-0.1 x 1.667 = -0.1667
4	1	-0.667	0.0667
5	0	0.333	-0.333

   c. Calculating the probability of iteration 1: $P_{ik}{}^{(1)} = \frac{e_{ik}{}^{(1)}}{\sum_{j=0}^{2} e_{ij}{}^{(1)}}$

   For i = 1:

$$
\begin{gathered}
e^{s^{(10)}} = e^{-0.0333} = 0.9673 \\
e^{s^{(11)}} = e^{0.0667} = 1.069 \\
e^{s^{(12)}} = e^{0.9673} = 0.8464 \\
\sum = 0.9673 + 1.069 + 0.8464 = 2.8827 \\
P_{10}{}^{(1)} = \frac{0.9673}{2.8827} = 0.3356 \\
P_{11}{}^{(1)} = \frac{1.069}{2.8827} = 0.3708 \\
P_{10}{}^{(1)} = \frac{0.8964}{2.8827} = 0.2936
\end{gathered} \tag{5}
$$

After updating the logit scores in the first iteration, the learning process continued to the second iteration by applying the same residual formula: $g^{(1)}{}_{ik} = P^{(1)}{}_{ik} - 1_{[y1=k]}$. These residuals were then used to update the logit scores using the equation $S^{(2)}{}_{ik} = S^{(1)}{}_{ik} - \eta\, g^{(1)}{}_{ik} = -0.1 \text{ x } g^{(1)}{}_{ik}$. In the second iteration, the logit for the target class increased—for example, from 0.0667 to 0.1311, and the corresponding probability rose from approximately 0.3708 to 0.3884. This pattern continued in the third iteration, where the residual for the target class decreased to −0.6116, the logit rose to 0.1933, and the predicted probability for the correct class reached 0.4098.

These changes indicate that the model progressively adjusted its predictions based on previous errors, with the class probability distributions gradually aligning more closely to the true labels.

During model training with the CatBoost algorithm, iterative updates are performed based on residuals—the difference between actual targets and predicted probabilities. Each iteration adds a weak learner to improve prediction accuracy and update class probabilities. Manual calculations up to the third iteration show that prediction probabilities begin to align with the actual targets. This is indicated by decreasing residual values, which approach zero in most cases, and a consistent reduction in the loss function. These results suggest that the model is converging, with notable improvement by the third iteration. The trained model is then used to predict the test data. The following is an illustration of the calculation on 10 test data samples.

Table 8. Test Data

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	y
3	1	2	3	2	2	2	1	1	3	1	1	3	2	0
3	2	2	2	2	1	3	2	2	3	3	3	3	4	1
3	2	3	2	1	3	2	4	1	3	4	2	4	4	0
2	2	1	2	2	1	2	3	2	3	2	1	3	3	0
2	2	3	1	3	2	2	2	4	1	1	4	2	3	1

1. Logit formula per iteration: $logit_k{}^{(t)} = logit_k{}^{(t-1)} + \eta \text{ x } w_k{}^{(t)}$
   Explanation:
   $Logit_k^{(t)}$: updated logit for class k at iteration t.
   η : learning rate.
   $w_k{}^{(t)}$: prediction from weak learner at iteration t
   a. Iteration 1: $logit_{k0}{}^{(1)} = 0 + 0.1 \text{ x } 0.3 = 0.03$; $logit_{k1}{}^{(2)} = 0 + 0.1 \text{ x } 0.0 = 0.00$; $logit_{k1}{}^{(2)} = 0 + 0.1 \text{ x}(-0.3) = -0.03$.
   b. Iteration 2: $logit_{k0}{}^{(2)} = 0.03 + 0.1 \text{ x } (-0.1) = 0.03 - 0.01 = 0.02$; $logit_{k1}{}^{(2)} = 0 + 0.1 \text{ x } 0.4 = 0.04$; $logit_{k1}{}^{(2)} = 0.03 + 0.1 \text{ x}(-0.3) = -0.03 - 0.03 = -0.06$.
   c. Iteration 3: $logit_{k0}{}^{(3)} = 0.02 + 0.1 \text{ x } 0.1 = 0.02 + 0.01 = 0.03$; $logit_{k1}{}^{(3)} = 0.04 + 0.1 \text{ x}(-0.2) = 0.04 - 0.02 = 0.02$; $logit_{k1}{}^{(3)} = -0.06 + 0.1 \text{ x } 0.3 = -0.06 + 0.03 = -0.03$.
2. The logit can be transformed into a probability using the following formulation:

$$
\begin{gathered}
P(k)\frac{e^{logit_k}}{\sum_j^3 = 1\ e^{logit_j}} \\
e^{0.03} = 1.0305 \\
e^{0.02} = 1.0202 \\
e^{-0.03} = 0.9704 \\
\text{Total} = 1.0305 + 1.0202 + 0.9704 = 3.0211 \\
P(K0) = \frac{1.0305}{3.0211} = 0.3412 \\
P(K1) = \frac{1.0202}{3.0211} = 0.3377 \\
P(K2) = \frac{0.9704}{3.0211} = 0.3211
\end{gathered} \tag{6}
$$

In the CatBoost algorithm, each iteration updates logits for each class using learned values (e.g., 0.3, 0.2, −0.3) based on gradient errors. These accumulated logits are then converted into class probabilities using the softmax function. $P(K0)$ is the predicted probability that the input belongs to class K0; a value of 0.3412 means a 34.12% likelihood, derived from the accumulated logits over iterations.

Feature importance analysis from the CatBoost model identified Q14, Q5, Q2, and Q8 as the top stress predictors. Q14 reflects academic difficulty, Q5 links procrastination to stress, Q2 emphasizes sleep quality, and Q8 highlights poor time management. In contrast, Q3 (diet) and Q4 (physical activity) had minimal impact. These results suggest targeted interventions such as lifestyle counselling, time management, and sleep improvement programs in campus mental health services.

Table 9. Importance of Features

Rating	Question	Level of Influence
1	Q14	9.75
2	Q5	9.42

3	Q2	8.96
4	Q8	8.95
5	Q9	8.10
6	Q12	7.26
7	Q10	6.94
8	Q6	6.81
9	Q7	6.49
10	Q1	6.01
11	Q11	5.95
12	Q13	5.64
13	Q4	6.06
14	Q3	4.65

3.4 Prediction Model Performance Evaluation

The implementation of the CatBoost model is done using Python on the Google Colab platform. The stages carried out include calling the library, uploading and reading data, identifying question columns, calculating total scores, creating stress level labels, separating training data and test data, training the CatBoost model, predicting results, and visualization. The results of predicting student stress levels by the CatBoost model on the whole data are presented in Figure 4. The predicted stress levels are categorized into Low, Medium, and High. Each category is labeled with the number of respondents to clearly show the distribution across classes.

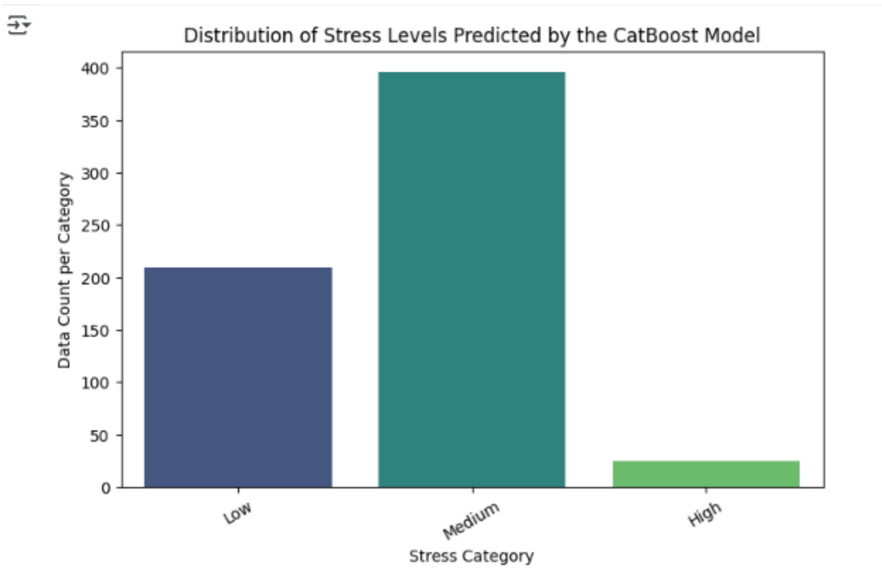


Figure 4. Distribution of Stress Levels Predicted by the CatBoost Model

The prediction results show that the majority of students are predicted to be in the "Medium" stress level category, followed by the "Low" category, and the least in the "High" category.

The testing phase was conducted to evaluate the model's performance in predicting student stress levels based on test data. The metrics calculated are *accuracy, precision, recall,* and *F1-score.*

1. Accuracy

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of data}} = \frac{1+29+77}{126} = \frac{107}{126} = 0.8492 = 0.85 \quad (7)$$

2. Precision (Per Class)
   a. Low: TP=1, FP=0 (medium to low) = 1

$$Precision = \frac{1}{1+0} = \frac{1}{1} = 1.00 \quad (8)$$

   b. Medium: TP = 77, FP = 6 (low to medium)

$$Precision = \frac{77}{77+8} = \frac{77}{85} = 0.906 \quad (9)$$

   c. High: TP = 29, FP = 2 (medium to high)

$$Precision = \frac{29}{29+2} = \frac{29}{31} = 0.935 \quad (10)$$

3. Recall (per class)
   a. Low: TP = 29, FN =13

$$Recall = \frac{29}{29+13} = \frac{29}{42} = 0.690 \tag{11}$$

b. Medium: TP = 77, FN = 2

$$Recall = \frac{77}{77+2} = \frac{77}{79} = 0.975 \tag{12}$$

c. High: TP = 1, FN = 4

$$Recall = \frac{1}{1+4} = \frac{1}{5} = 0.20 \tag{13}$$

4. F1 Score

$$F1\text{ Low} = 2 \text{ x } \frac{0.935 \text{ x } 0.890}{0.935+0.890} = 0.794 \tag{14}$$

$$F1\text{ Medium} = 2 \text{ x } \frac{0.906 \text{ x } 0.975}{0.906+0.975} = 0.94 \tag{15}$$

$$F1\text{ High} = 2 \text{ x } \frac{1.00 \text{ x } 0.20}{1.00+0.20} = 0.33 \tag{16}$$

Based on the evaluation results presented in the classification report, the model achieved an accuracy of 85.00%. The model demonstrated strong performance in classifying the medium stress category, with a precision of 0.82, recall of 0.97, and F1-score of 0.89. For the low and high stress categories, the model produced varied results, with F1-scores of 0.79 and 0.33 respectively. This indicates that the model is highly effective at predicting medium stress, reasonably effective for low stress, but still shows weaknesses in predicting high stress.

These results are supported by the confusion matrix (Figure 4), which shows that for the medium stress category, the model correctly classified 77 out of 79 test samples, resulting in a precision of 0.82, recall of 0.97, and F1-score of 0.89. For the low stress category, 29 out of 42 samples were correctly classified, while the rest were misclassified as medium stress, yielding a precision of 0.94, recall of 0.69, and F1-score of 0.79. In the high stress category, only 1 out of 5 samples was correctly classified, with the remaining 4 misclassified as medium stress, producing a perfect precision of 1.00, but a low recall of 0.20 and F1-score of 0.33.

```
Classification Report:
              precision    recall  f1-score   support

        High       1.00      0.20      0.33         5
         Low       0.94      0.69      0.79        42
      Medium       0.82      0.97      0.89        79

    accuracy                           0.85       126
   macro avg       0.92      0.62      0.67       126
weighted avg       0.87      0.85      0.84       126


Accuracy Score:  0.8492063492063492
```

Figure 5. Classification Report

The confusion matrix shows that misclassifications only occurred between adjacent stress levels, with no direct errors between low and high categories. This indicates the model effectively distinguished all stress levels, especially the medium class. One key challenge in stress-level classification is class imbalance, where the medium category dominates and overshadows low and high levels. To address this, the dataset was balanced using SMOTE, ensuring equal samples per class. This preprocessing step significantly improved the model's ability to recognize all stress categories, including the extreme ones.

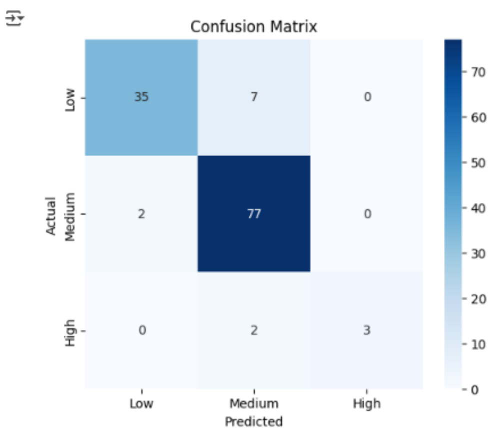


Figure 6. Confusion Matrix

The implementation of the CatBoost algorithm on student questionnaire data yielded an overall accuracy of 84.92%. Beyond accuracy, the model demonstrated strong and consistent performance across all three stress levels—Low, Medium, and High—with relatively balanced precision, recall, and F1-scores. The classification report (Figure 4) and confusion matrix (Figure 5) confirmed that the model predicted medium stress with high accuracy, while the performance on extreme classes was lower, particularly on the High stress class. This performance was supported by data balancing using SMOTE, well-curated lifestyle-based features, and a structured preprocessing pipeline.

In comparison, a previous study by Astari et al [33] implemented the Random Forest algorithm to classify stress levels during sleep and reported a higher accuracy of 93.65%. However, their study focused solely on physiological signals such as snoring frequency, breathing rate, body temperature, and heart rate, with no consideration of class balancing techniques or reporting of metrics like F1-score and recall per class. Additionally, their paper does not provide analysis of classification errors or confusion matrices, which limits interpretability and generalization. Unlike that approach, this study used lifestyle-based questionnaire data factors that are easier to collect, more actionable for stress management interventions, and more interpretable for educational settings. Moreover, evaluation was carried out not only through overall accuracy but also via a complete breakdown of performance per class, making the results more robust and informative. While Random Forest remains a strong baseline in classification tasks, the use of CatBoost in this study provides notable advantages, particularly in handling categorical features, resistance to overfitting, and interpretability of model behaviour. These considerations reinforce the reliability of CatBoost as an optimal algorithm in this context, supported by balanced data and transparent performance analysis.

4. CONCLUSION

This study successfully applied the CatBoost algorithm to classify student stress levels based on 630 questionnaire responses from the Faculty of Science and Technology. The model achieved an accuracy of 91.27% and effectively identified patterns between lifestyle factors and stress levels. Results showed that 167 students experienced mild stress, 317 moderate, and 20 severe. The most influential factors were related to time management, particularly in organizing daily tasks, followed by sleep quality and physical activity—specifically, task prioritization (Q5) and bedtime consistency (Q2) emerged as key indicators associated with elevated stress levels.

This study contributes to the growing body of literature on mental health prediction by demonstrating the feasibility of using CatBoost in handling categorical lifestyle data with high interpretability and robust accuracy. Compared to other ensemble models, CatBoost also offers superior handling of categorical features and resistance to overfitting. Performance was further validated using precision, recall, and F1-score metrics, ensuring balanced classification across stress levels, as shown in the confusion matrix.

Despite promising results, this study is limited to a single faculty, which may reduce the generalizability of the findings. Additionally, the use of purposive sampling within a single academic domain may introduce selection bias, which should be addressed in future research through multi-centere sampling strategies. The model also holds potential for practical implementation in student support services, enabling early identification and targeted interventions. Future work may explore comparison with other algorithms such as XGBoost and SVM to determine the most effective and interpretable approach for broader populations.

5. REFERENCES

[1] M. Anastasia, V. S. Maulivia, and S. Suharjito, "Machine learning methods for assessing depression and mental health data," *INTECOMS J. Inf. Technol. Comput. Sci.*, vol. 7, no. 3, pp. 606–612, 2024, doi: 10.31539/intecoms.v7i3.9584.

[2] Z. Mufatihah, A. P. Zelya, R. A. Puriani, and R. M. Putri, "The phenomenon of academic stress among university students," *J. Educ. Res.*, vol. 6, no. 1, pp. 573–580, 2025, doi: 10.47827/jer.v6i1.564.

[3] K. A. S. Widiasa, D. N. Adiwibawa, I. Syuhada, and I. Winangun, "The relationship between stress levels in thesis writing and sleep quality among FK Unizar students batch 2019," *Cakrawala Med. J. Health Sci.*, vol. 2, no. 1, pp. 68–76, 2023, doi: 10.59981/5hergy11.

[4] World Health Organization (WHO), *World mental health report: Transforming mental health for all.* Geneva: WHO, 2022. [Online]. Available: https://www.who.int/publications/i/item/9789240049338

[5] R. K. Djoar and A. P. M. Anggarani, "Factors affecting academic stress among final-year students," *Jambura Health Sport J.*, vol. 6, no. 1, pp. 52–59, 2024, doi: 10.37311/jhsj.v6i1.24064.

[6] A. P. Abdullah, I. Royani, P. N. Hamzah, S. Amir, and M. A. Jaya, "The relationship between academic workload and stress levels among UMI Medical Faculty students batch 2020," *Fakumi Med. J.*, vol. 4, pp. 110–119, 2024, doi: 10.33096/fmj.v4i2.393.

[7] S. R. Anugrahaini, M. Fatchan, and T. N. Wiyatno, "Stress level prediction models among students using artificial neural network algorithms with Orange," *J. Inf. Eng.*, vol. 9, pp. 1–16, 2025, doi: 10.36040/jati.v9i3.13759.

[8] N. M. Yusuf and J. M. Yusuf, "Factors influencing academic stress," *Psyche 165 J.*, vol. 13, pp. 235–239, 2020, doi: 10.35134/jpsy165.v13i2.84.

[9] S. Anisa, A. Komarudin, and E. Ramadhan, "Classification system to determine general student stress levels using K-Nearest Neighbors method," *J. Inf. Technol. Sci.*, vol. 6, pp. 568–578, 2024, doi: 10.51401/jinteks.v6i3.4317.

[10] R. Qasrawi *et al.*, "Assessment and prediction of depression and anxiety risk factors in schoolchildren: Machine learning techniques performance analysis," *JMIR Form. Res.*, vol. 6, no. 8, pp. 1–15, 2022, doi: 10.2196/32736.

[11] S. B. Seto, M. T. S. Wondo, and M. F. Mei, "The relationship between motivation and student stress levels in writing undergraduate theses," *J. Basicedu*, vol. 4, no. 3, pp. 733–739, 2020, doi: 10.31004/basicedu.v4i3.431.

[12] R. Tasalim and A. R. Cahyani, *Academic Stress and Its Management*, Banjarmasin: Guepedia, 2021. ISBN: 978-623-270-839-6. [Online]. Available: www.guepedia.com.

[13] N. Widyawati, M. Khasanah, Muttaqin, E. Rasywir, and A. Feranika, "Stress level prediction among students of Universitas Dinamika Bangsa Jambi during hybrid learning using Naive Bayes algorithm," *J. Inf. Syst. Manag.*, vol. 2, pp. 99–110, 2022, doi: 10.33998/jms.2022.2.1.44.

[14] W. Clariska, Yuliana, and Kamariyah, "The relationship between stress level and sleep quality among final-year students of the Faculty of Medicine and Health Sciences, Universitas Jambi," *J. Indones. Nurs. Sci.*, vol. 1, pp. 94–102, 2020, doi: 10.22437/jini.v1i2.13516.

[15] Hariaty, V. Elita, and A. Dilaluri, "Overview of stress among final-year students working on their theses," *J. Prof. Nurs.*, vol. **11, pp. 1–7, 2023, doi: 10.33650/jkp.v11i1.5539.**

[16] M. Maulidah and N. Hidayati, "Prediction of sleep health and lifestyle using machine learning," *Comput. Netw. Technol.*, vol. 4, pp. 81–86, 2024, doi: 10.31294/conten.v4i1.4918.

[17] M. Saifudin, S. R. Adawiyah, and I. Mukhaira, "Factors influencing academic stress levels among non-regular nursing students," *J. Health*, vol. 12, no. 2, pp. 199–207, 2023, doi: 10.37048/kesehatan.v12i2.267.

[18] Y. Suharsono and Z. Anwar, "Stress analysis and self-adjustment among university students," *J. Online Psychol.*, vol. 8, pp. 1–12, 2020, doi: 10.22219/cognicia.v8i1.11527.

[19] N. Bachtiar, M. I. Hasan, and C. Andodo, "Academic stress among health students related to psychological, physiological changes, motivation, and learning quality," *J. Health Sci. Technol.*, vol. 14, pp. 11–20, 2023, doi: 10.36308/jik.v14i2.539.

[20] N. D. Syandika and W. Yustanti, "Anomaly detection in transaction cancellation on Tiktok Shop using CatBoost algorithm," *J. Inf. Comput. Sci.*, vol. 5, pp. 149–156, 2023, doi: 10.26740/jinacs.v5n02.p149-156.

[21] Z. Fan, J. Gou, and S. Weng, "A feature importance-based multi-layer CatBoost for student performance prediction," *IEEE Trans. Knowl. Data Eng.*, vol. 36, no. 11, pp. 5495–5507, 2024, doi: 10.1109/TKDE.2024.3393472.

[22] R. Irfannandhy, L. B. Handoko, and N. Ariyanto, "Performance analysis of Random Forest and CatBoost with SMOTE for diabetes risk prediction," *Edumatic J. Inf. Educ.*, vol. 8, pp. 714–723, 2024, doi: 10.29408/edumatic.v8i2.27990.

[23] T. Z. Jasman, M. A. Fadhullah, A. L. Pratama, and Rismayani, "Analysis of Gradient Boosting, AdaBoost, and CatBoost algorithms in water quality classification," *J. Inf. Syst. Eng.*, vol. 8, pp. 392–402, 2022, doi: 10.28932/jutisi.v8i2.4906.

[24] D. M. Putri, Fahrizal, A. Rahmawati, and B. W. Akramunnas, "Comparison of CatBoost and XGBoost algorithms in heart disease classification," *J. Appl. Technol.*, vol. 15, pp. 126–133, 2023, doi: 10.30606/aptek.v15i2.1930.

[25] G. Ravindiran *et al.*, "Air quality prediction by machine learning models: A predictive study on the Indian coastal city of Visakhapatnam," *Chemosphere*, vol. 338, 2023, doi: 10.1016/j.chemosphere.2023.139518.

[26] A. M. E. F. Firmandani, A. Hudawi, and A. Tholib, "CatBoost model optimization with feature selection and hyperparameter tuning for potential bank customer prediction," *Acad. J. Comput. Sci. Res.*, vol. 6, pp. 118–127, 2024, doi: 10.38101/ajcsr.v6i2.15656.

[27] A. F. Istianto, A. Id Hadiana, and F. R. Umbara, "Rainfall prediction using Categorical Boosting (CatBoost) method," *J. Inf. Eng. Student J.*, vol. 7, no. 4, pp. 2930–2937, 2023, doi: 10.36040/jati.v7i4.7304.

[28] H. Mulyo and A. K. Zyen, "The effect of hyperparameter tuning on Gradient Boosting for predicting new student major selection," *Bull. Comput. Sci. Res.*, vol. 5, no. 2, pp. 131–137, 2025, doi: 10.47065/bulletincsr.v5i2.454.

[29] W. Nugraha and M. Syarif, "Weighting techniques to address class imbalance in churn prediction using XGBoost, LightGBM, and CatBoost," *Techo.Com*, vol. 22, pp. 97–108, 2023, doi: 10.33633/tc.v22i1.7191.

[30] M. T. Syamkalla, S. Khomsah, and Y. S. R. Nur, "CatBoost and SHAP implementation for predicting indie game popularity on Steam," *J. Inf. Technol. Comput. Sci.*, vol. 11, no. 4, pp. 777–786, 2024, doi: 10.25126/jtiik.1148503.

[31] A. Ulul, A. Wafiqi, B. A. James, A. H. Ramadhan, and A. Nizar, "Stress level prediction in students of UNUGHA Cilacap," *J. Teknokompak*, vol. 18, no. 2, pp. 331–343, 2024, doi: 10.33365/jtk.v18i2.3933.

[32] F. N. Zahrah and Muljono, "Machine learning for student stress detection: Perceptron as an effective classification model for early intervention," *Edumatic J. Inf. Educ.*, vol. 8, no. 2, pp. 764–773, 2024, doi: 10.29408/edumatic.v8i2.28011.

[33] D. F. Astari, Y. H. Chrisnanto, and Melina, "Classification of stress levels during sleep using Random Forest algorithm," *J. Inf. Eng.*, vol. 7, no. 5, pp. 3676–3684, 2023, doi: 10.36040/jati.v7i5.7750.

[34] D. Kurniawan, *Introduction to Machine Learning with Python*, Jakarta: Elex Media Komputindo, 2021. ISBN: 978-623-00-2174-9.