Uluslararası İleri Doğa Bilimleri ve Mühendislik Araştırmaları Dergisi
Sayı 9, S. 210-214, 10, 2025

Araştırma Makalesi

IJANSER
https://as-proceeding.com/index.php/ijanser
ISSN:2980-0811

International Journal of Advanced Natural Sciences and Engineering Researches
Volume 9, pp. 210-214, 10, 2025

Research Article

Classification of Student Stress Levels Using Machine Learning Methods

Kadircan Yeler[*,1] and Selim Sürücü [2]

[1] *Bilgisayar Mühendisliği Bölümü, Çankırı Karatekin Üniversitesi, Türkiye.*
[2] *Bilgisayar Mühendisliği Bölümü, Çankırı Karatekin Üniversitesi, Türkiye Orcid: 0000-0002-8754-3846*

[*] *(yelerkadircan@gmail.com) Email of the corresponding author*

(Received: 17 October 2025, Accepted: 21 October 2025)

(5th International Conference on Trends in Advanced Research ICTAR 2025, October 16-17, 2025)

ATIF/REFERENCE: Yeler, K. & Sürücü, S. (2025). Classification of Student Stress Levels Using Machine Learning Methods, *International Journal of Advanced Natural Sciences and Engineering Researches*, 9(10), 210-214.

Abstract – Nowadays, students feel under pressure and experience stress for various reasons. These reasons generally manifest as academic pressures, the social environment, and personal anxieties. Intense stress is a significant factor that negatively impacts students' academic success, psychological health, and overall quality of life. The inadequacy of traditional subjective assessment methods in determining stress levels accurately and reliably has increased the need for objective, data-driven solutions. The main purpose of this study is to automatically classify student stress levels (Low, Medium, High) using machine learning algorithms and the Student Stress Monitoring Dataset from Kaggle. This dataset contains 1,100 observations, 21 features, and no missing values. In this study, Logistic Regression, Random Forest, and XGBoost models were applied for classification. The accuracy of these models was measured as 88.1%, 86.2%, and 86.8%, respectively. The results show that machine learning methods can be used effectively in predicting student stress levels.

Keywords – Stress, Student stress prediction, Student Stress Monitoring Dataset, Logistic Regression, Random Forest, XGBoost.

I. INTRODUCTION

Students are subjected to chronic stress due to the high expectations of modern academic life and families, heavy course loads, the necessity of adapting to a competitive career environment, and the dynamics of complex social lives [1,2]. Stress is considered a significant public health issue that not only reduces students' academic performance and negatively impacts their concentration and memory functions, but also directly threatens their psychological well-being and overall quality of life, leading to issues such as anxiety and depression [3]. This situation challenges educational institutions to identify at-risk students early and rapidly and to develop preventative intervention strategies.

The sources of student stress are not limited to emotional or cognitive factors. In addition to these factors, stress sources can be grouped under three main headings: psychological, physiological, and environmental [4,5]. These factors include:

- Psychological Factors: Variables such as test anxiety, self-esteem, and depression reflect students' mental state. These variables have a significant impact on stress levels.
- Physiological Factors: The physical effects of stress, such as insomnia, headache, and high blood pressure, appear as direct biological indicators.

- Environmental Factors: External factors such as academic load, living conditions, noise level and academic performance expectations have a significant impact on increased stress levels [6].

Traditional methods for assessing stress levels typically rely on surveys and subjective assessments based on students' own perceptions. However, these subjective approaches may fall short of accurately and reliably determining stress levels due to factors such as individual differences, cultural factors, and the manipulation of responses [7].

As a solution to this shortcoming, data-driven approaches, particularly Machine Learning (ML) methods, have seen significant development in recent years. ML algorithms can analyze complex patterns and interactions between factors in multidimensional datasets (psychological, physiological, and environmental) [8]. These analyses can address classification features overlooked by traditional methods and clearly reveal their effects on stress levels [9].

A review of the literature reveals studies specifically utilizing features derived from survey data to classify stress levels using machine learning algorithms [10]. This study, to contribute to this field, employed a comprehensive student stress dataset shared on the Kaggle platform. The aim was to compare the performance of different classification algorithms on this dataset. The study's unique contribution is the rigorous comparison of the performance of three powerful classifiers with different architectures (Logistic Regression, Random Forest, XGBoost) on a high-dimensional dataset. Furthermore, it aims to uncover the most suitable, reliable, and practical model for automatically predicting student stress levels. This provides a scientific and applicable approach for many institutions and organizations, especially educational institutions, to identify student stress levels earlier and develop targeted intervention mechanisms.

II. Materials and Method

A. Dataset

The Student Stress Monitoring Dataset used in this study is shared on the Kaggle platform [11]. This dataset contains various psychological, physiological, and environmental factors that affect students' stress levels.

- **Size**: The dataset consists of observation data from a total of 1,100 students.
- **Features**: It contains a total of 21 independent features and a target variable (stress level).
- **Target Variable (stress_level):** This dependent variable, representing students' stress levels, is classified into three categories: 0 (Low Stress), 1 (Medium Stress), and 2 (High Stress). Having three different categories presents a multi-class classification problem.
- **Data Balance**: The dataset has a balanced class distribution, with no significant imbalance between low, medium, and high stress classes, which reduces class bias in model training. The distribution of Stress Level classes is shown in Fig. 1.

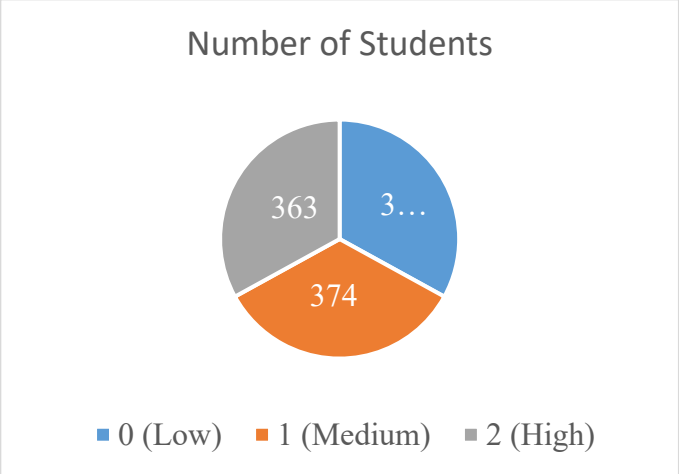

Fig. 1 Number of students according to stress levels

B. *Pre-processing Stages*

Pre-processing steps were performed on the dataset to ensure the accurate and efficient operation of the machine learning models. These pre-processing steps are as follows:

- **Missing Value Checking and Data Cleaning:** The dataset was initially determined to contain no missing values (using the df.isnull(). sum () check). Therefore, no row deletion or missing value filling was performed. Therefore, all 1,100 observations in the dataset were used in this study.
- **Feature Separation:** The target variable (stress_level) was selected as the dependent variable (y). The other 20 features were separated as independent variables (x).
- **Feature Scaling (Standardization):** Because the different scales of independent variables can negatively impact the performance of distance-based and scale-sensitive algorithms, especially Logistic Regression, all features were standardized using StandardScaler (mean 0, standard deviation 1).
- **Training and Test Data Separation:** To assess the generalization ability of the models, the dataset was randomly split into two parts: 80% training and 20% test, using the train_test_split function. This means 220 observations are reserved for testing.

C. *Models*

Three different machine learning models were used for the classification using the dataset above. These models are:

- **Logistic Regression:** This is a linear classifier with high interpretability used for categorical dependent variables [12, 13]. In this study, we used its extended version (0, 1, 2) to accommodate the multi-class output variable.
- **Random Forest:** This is a powerful model based on decision trees, reducing the risk of overfitting and offering high accuracy. It was chosen for this study due to its ability to capture complex and non-linear relationships [14].
- **XGBoost (Extreme Gradient Boosting):** This boosting model is known for its superior performance on structured data and is optimized for speed and scalability. It was chosen for its ability to build a strong model by sequentially training weak learners and correcting previous errors.

III. RESULTS

Each implemented classification model was evaluated on the pre-processed test dataset (20%). There are a total of 220 observations in the test set (Class 0: 76, Class 1: 73, Class 2: 71 support observations). Accuracy was used as the main performance metric, and the ability of the models to distinguish each stress level (Class 0, 1, 2) was additionally analyzed using Confusion Matrix analysis. The accuracy value of the performance of the models on the test data is shown in Table 1.

Table 1. Performance results of the models on the test data (according to the accuracy metric)

Model	Accuracy (%)
Logistic Regression	88.1
Random Forest	86.2
XGBoost	86.8

When Table 1 is examined, it is observed that the Logistic Regression model showed the best performance with an accuracy rate of 88.1%, which is slightly higher than the other two ensemble algorithms. Random Forest and XGBoost models, on the other hand, exhibited very close and high performances. These results confirm that all three models were quite successful in classifying the dataset.

The performance of the models on each class was examined in more detail using the Precision, Recall, and F1-Score metrics. The results are shown in Figure 2. Logistic Regression achieved the highest Precision (0.91) and F1-Score (0.91) values for predicting high stress levels (Class 2), demonstrating that it was able

to distinguish high stress conditions more reliably than the other models. Thanks to the balanced class distribution, all models did not exhibit significant imbalances in performance between classes.

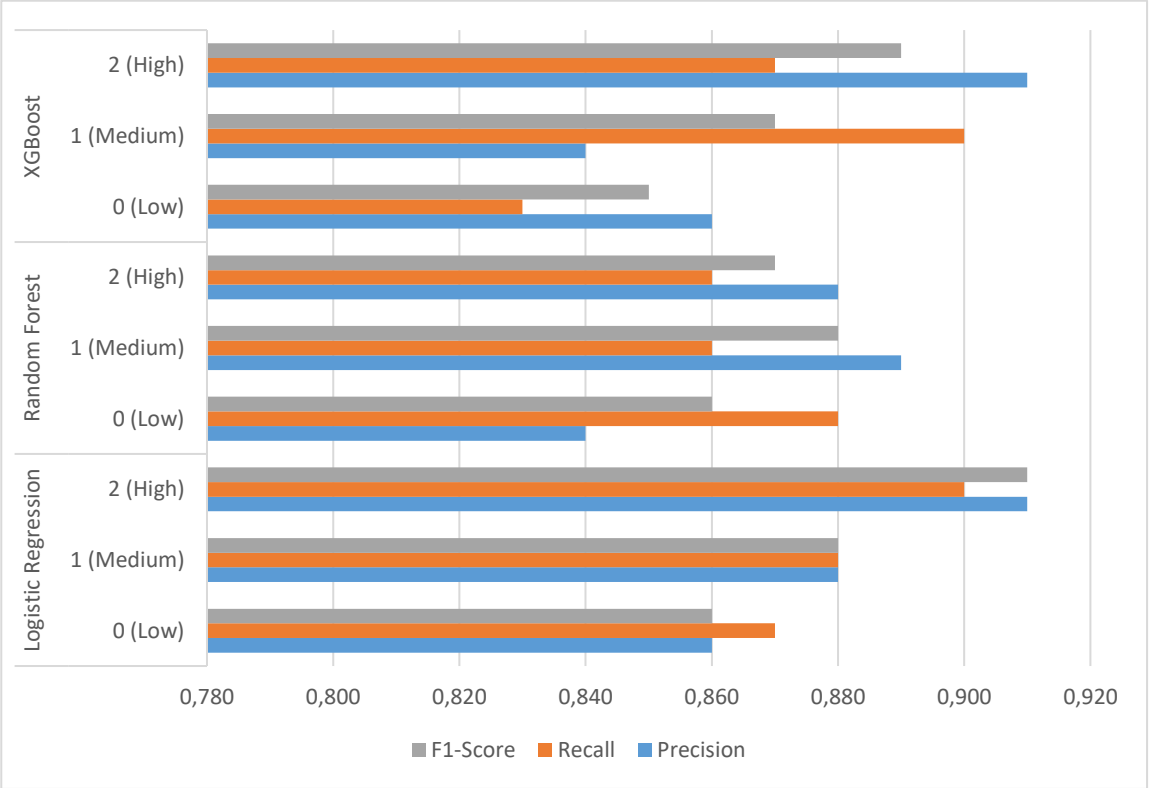

Fig. 2 Number of students according to stress levels

IV. DISCUSSION

Experimental results definitively demonstrate that student stress levels can be classified with high accuracy (up to 88.1%) using machine learning algorithms. The most striking finding is that the Logistic Regression model slightly outperforms more complex and advanced ensemble algorithms such as Random Forest and XGBoost in terms of accuracy. This provides significant evidence that, given the standardized features and balanced class structure of the dataset, there is a strong linearly separable structure among the variables determining stress levels. In other words, even a relatively simple linear classifier (Logistic Regression) was found to effectively model the relationships in this 20-feature dataset.

REFERENCES

[1] G. A. Gobena, "Effects of Academic Stress on Students' Academic Achievements and Its Implications for Their Future Lives," Anatolian Journal of Education, c. 9, sy. 1, ss. 113-130, 2024.

[2] M. C. Pascoe, S. E. Hetrick and A. G. Parker, "The impact of stress on students in secondary school and higher education," *International Journal of Adolescence and Youth*, vol. 25, no. 1, pp. 104-112, 2020.

[3] L. Rosenthal, S. Lee, P. Jenkins, J. Arbet, S. Carrington, S. Hoon, S. K. Purcell and P. Nodine, "A Survey of Mental Health in Graduate Nursing Students during the COVID-19 Pandemic," *Nurse Educ.*, c. 46, sy. 4, ss. 215-220, 2021.

[4] D. Bedewy and A. Gabriel, "Examining perceptions of academic stress and its sources among university students: The Perception of Academic Stress Scale," Health Psychology Open, vol. 2, no. 2, pp. 1-9, 2015.

[5] M. G. I. Emran, S. Mahmud, A. H. Khan, N. N. Bristy, A. K. Das, R. Barma, A. Barma, M. H. Mita, L. Bosunia, M. Rahman and M. Roy, "Factors Influencing Stress Levels Among Students: A Virtual Exploration," European Journal of Medical and Health Sciences, c. 6, sy. 6, ss. 1-13, 2024.

[6] N. Najafi, K. Movahed, Z. Barzegar and S. Samani, "Environmental Factors Affecting Students' Stress in the Educational Environment: A Case Study of Shiraz Schools," International Journal of School Health, c. 5, sy. 2, ss. 1-7, 2018.

[7] S. Arya, A. Anju and N. A. Ramli, "Predicting the stress level of students using Supervised Machine Learning and Artificial Neural Network (ANN)," Indian Journal of Engineering, c. 21, ss. 1-24, 2024.

[8] E. Abdelfattah, S. Joshi and S. Tiwari, "Machine and Deep Learning Models for Stress Detection Using Multimodal Physiological Data," IEEE Access, vol. 13, pp. 4597-4608, 2025.

[9] E. Lazarou and T. P. Exarchos, "Predicting stress levels using physiological data: Real-time stress prediction models utilizing wearable devices," AIMS Neuroscience, vol. 11, no. 2, pp. 76-102, 2024.
[10] 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, c. 13, sy. 22, ss. 3455, 2023.
[11] M. S. I. Ovi, J. Hossain, M. R. A. Rahi and F. Akter, "Protecting Student Mental Health with a Context-Aware Machine Learning Framework for Stress Monitoring," arXiv preprint, arXiv:2508.01105, 2025.
[12] M. P. LaValley, "Logistic regression," Circulation, vol. 117, no. 18, pp. 2395-2399, 2008.
[13] S. Çelebioğlu and S. Sürücü, "Predicting the University Placement Status of University Students Using Artificial Intelligence," International Journal of Advanced Natural Sciences and Engineering Research, c. 7, sy. 2, ss. 1-4, 2023.
[14] P. Probst, N. W. Marvin and A. Boulesteix, "Hyperparameters and tuning strategies for random forest," Wiley Interdisciplinary Reviews: data mining and knowledge discovery, vol. 9, sy. 3, 2019.
[15] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16), San Francisco, CA, USA, Aug. 2016, pp. 785–794.