

Student Stress Level Prediction Using Machine Learning

Pallavi, Dr. Vishal Verma, Sonu

MCA Student, Professor, Asst. Professor

Deptt. of Comp. Sc. and Application, Ch. Ranbir Singh University, Jind, Haryana, India

Abstract: Student stress is a vital problem in educational institutions as it impacts the academic achievements of students, their mental condition, and general well-being. The early recognition of stress may aid students in getting the proper support. The aim of this research is to predict the stress level of students through machine learning approaches. The data used in this study is provided by Kaggle and comprises such variables as Sleep Quality, Headache, Academic Performance, Study Load, and Extracurricular Activities. Two supervised machine learning methods, Logistic Regression and Random Forest, were used for classification of student stress level. For the assessment of the models' performance, Accuracy, Precision, Recall, and F1-Score were used. As a result, Random Forest provided better performance and prediction accuracy than Logistic Regression. The results of this study show that machine learning can be efficiently used for prediction of student stress level.

Keywords: Student Stress Level Prediction, Machine Learning, Logistic Regression, Random Forest, Stress Level, Classification.

I. INTRODUCTION

The stress among students has become a major problem in the academic environment today. The students' mental and physical health is being affected by academic pressure, unhealthy lifestyle, competition and lack of proper balance between studies and personal life. Stress impacts not only academic performance but also concentration, confidence, emotional stability, and overall well-being [1].

The main purpose of this research paper is to predict the student stress levels using Machine Learning techniques. The study aims to analyze various factors affecting students' stress levels and to identify the patterns for early stress prediction. Machine Learning helps to understand the association of various lifestyle and academic factors with student stress[4].

The data set for this study includes some of the key characteristics of the students, including Sleep Quality, Headache Frequency, Academic Performance, Study Load, and Participation in Extracurricular Activities. These variables have been considered for this project because they have a significant effect on the life of the student as well as the mental illness of the student. The target variable of the dataset is the Stress Level, which is classified into various levels like low, moderate and high.[2]

As per the analysis, there are various factors that result in a high stress level among the students, which include poor sleep quality, high study load, frequent headaches, and low academic performance. Students who do not manage a balanced lifestyle amid the academic activities tend to suffer from higher stress levels. However, the involvement of extracurricular activities helps students decrease their stress levels and attain mental peace[2].

This study proves the use of machine learning in predicting the stress level of the student at an early stage.[3] Use of such prediction models will help in improving the well-being and performance of students. This makes it necessary for all educational institutions to think beyond their objective of ensuring academic excellence, and consider the mental health of their students[1].

The use of such a research also helps in making students aware of the significance of mental health in their lives. Use of technology together with education, through Machine Learning based stress prediction systems, can create a atmosphere.

Literature Review

There are many researchers who tried to predict student stress through the machine learning approach. In one such research conducted by Rahunathan et al. (2025), the use of two models is made for this purpose: namely, Naïve Bayes and KNN. Here the use of questionnaires like PSS and DASS-21 is made. According to this, it was seen that Naïve Bayes worked better than KNN, giving an accuracy of around 90% in one data set and 77% in another data set.

In another research study conducted by Dharaneedharan et al. (2025), the use of SVM, decision tree, ANN, and Adaboost models is made. In this case, according to their performance, decision tree is seen to give the best result. Next come ANN and AdaBoost followed by SVM which gives comparatively lower accuracy. It was also seen here that study time, sleep, and GPA play an important role in determining stress.

Another research by Mahalakshmi and Sujatha (2023) used models like KNN, Decision Tree, SVM, and Naïve Bayes. In their results, KNN gave the highest accuracy, around 87%. This shows that KNN works well for their dataset.

Another research by Disha Sharma, Dr. Sandeep Singh Kang(2020) used models like Logistic Regression, Random Forest, Bayes Net, Decision Tree, Naive Bayes but Random Forest gives the highest accuracy of 94.73%, making it the best model among all. It also shows that machine learning can help in early detection of stress, which can reduce serious health problems in students.

Another research by Rocco de Filippis, Abdullah AI Foysal(2024) used model Random Forest and other correlation analysis in this model Random Forest gives accuracy of 88% . it will help to identify different factors that cause stress in students and understand how these factors are connected with each other using machine learning techniques.

Another research by Dr. Anusuya Biswas (2024) used model FSS-LGBM (Fine-tuned Squirrel Search driven LightGBM). It is an advanced model combining optimization technique (FSS) with LightGBM. The proposed model FSS-LGBM performed better than other models like MLP, Deep Learning, and Decision Tree. It achieved around **98% accuracy**, which is very high. This shows that the model is very effective in predicting student stress and can help in early detection.

Another research by Shashank Shekhar and Abhay Kumar (2025) used models Random Forest, Decision Tree, XG Boost, SVM, KNN but all in these model Random Forest gives best accuracy among these model 65%. It will assist in predicting the stress level of the students through machine learning models.

3. Problem Statement

The stress levels of the students in educational institutions are determined by many educational and health factors. The data set used in this study is based on the questionnaires which include factors like Sleep Quality, Headache, Academic Performance, Study Load, and Extracurricular Activities. These factors help in predicting the stress level of the students. The machine learning algorithms Logistic Regression and Random Forest are used for classification of the stress level of the students. The aim of this study is to develop a model for prediction.

4. System Model

In this study, two supervised machine learning algorithms, namely logistic regression and random forest, are used to predict the stress levels of students. The system is developed using important factors such as Sleep Quality, Headache, Academic Performance, Study Load, and Extracurricular Activities. These factors are taken as input features, while the Stress Level is considered as the target variable.

The values of each feature are processed and analyzed by the machine learning models to identify the stress level of a student. Logistic Regression is used as a baseline classification model, whereas Random Forest is used as the main prediction model because of its ability to handle complex relationships among different factors. The stress levels predicted can thus be utilized in detecting students who are facing increased stress and therefore need immediate assistance and counseling.

5. Proposed Methodology

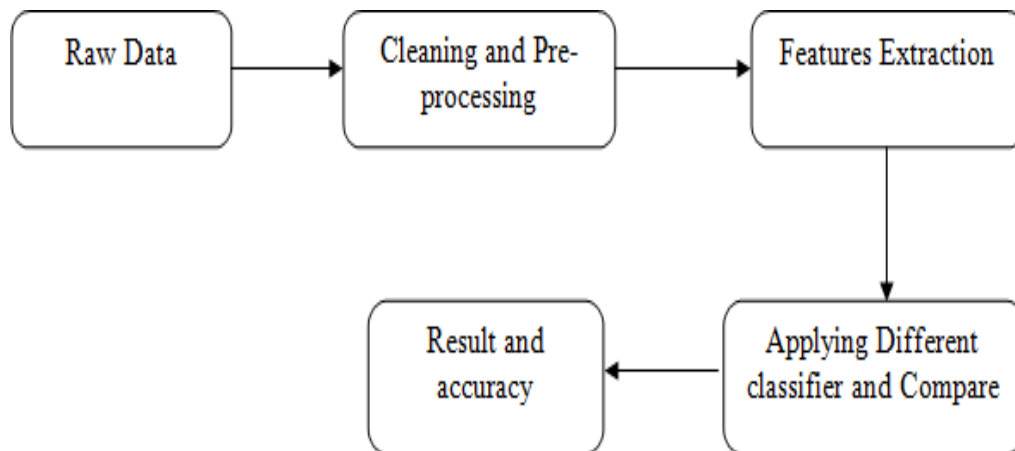


Fig. 5.1: Proposed Methodology

This is due to the fact that in the study, different machine learning techniques are used for predicting the stress level of Figstudents. In this regard, the study involves analyzing various variables influencing student stress such as Sleep Quality, Headache, Academic Performance, Study Load, and Extracurricular Activities. The architecture of the proposed system includes data collection from the Kaggle dataset, data preprocessing, data cleaning, feature selection, modeling, and evaluation. In this respect, Logistic Regression and Random Forest models are utilized for classifying the stress level of students, and then their performance is evaluated based on metrics.

1. DATA PREPARATION

The dataset used for this project was taken from Kaggle, and it is associated with the prediction of the stress levels of students. It is a dataset that includes data regarding some of the factors affecting the stress levels of students. This dataset has 520 instances of students and six attributes; out of these six attributes, five are input variables while one is the output variable.

The input variables include Sleep Quality, Headache, Academic Performance, Study Load, and Extracurricular Activities. These variables are different dimensions that can have an effect on the stress level of the students. The target variable is "How would you rate your stress levels?"

2. DATA DESCRIPTION

The dataset used in this study is a predefined dataset collected from Kaggle for predicting student stress levels using machine learning. It contains 520 student records, where each row represents information related to an individual student. The dataset mainly focuses on academic, health, and lifestyle factors that may affect a student's stress level.

This dataset includes 6 important attributes, which help in understanding the causes of student stress. These attributes are Sleep Quality, Headache, Academic Performance, Study Load, Extracurricular Activities, and Stress Level. Each feature is represented in numerical form, which makes the dataset easy to analyze and suitable for machine learning models.

Dataset Features S. No	Features	Type
1	Sleep Quality	Numeric
2	Headache	Numeric
3	Academic Performance	Numeric
4	Study Load	Numeric
5	Extracurricular Activities	Numeric
6	Stress Level	Numeric (Target Variable)

Table 5.1

we can easily predict the level of stress. In our dataset there are 5 level of stress and 1 output of predicting the stress of a student. As we discuss in introduction of there are level of stress which we are trying to predict using machine learning models like LR, , KNN, Decision Tree but here we only apply logistic regression and random forest.

3. Data Pre-processing

Before applying the machine learning algorithms, the dataset was carefully pre-processed to improve its quality and reliability. The dataset was checked for missing values, duplicate records, and any inconsistencies in the data. The missing values, if any, were taken care of in such a way that they did not impact the functioning of the prediction models.

All the features from the dataset, i.e., Sleep Quality, Headache, Academic Performance, Study Load, Extracurricular Activities, and Stress Level, have been quantified in the numerical format. Hence, there was no need for categorical encoding of the features in the pre-processing stage. The selection of input features and the output feature was done rightly after the data processing phase.

6. EXTRACTING TESTING AND TRAINING DATA .

The data set was split into two categories to help in developing and evaluating the models. About 80% of the data were used in training the machine learning models while 20% were used in testing the models. This is done to evaluate the performance of the machine learning models in predicting stress level values in unseen data.

7. ALGORITHMS

This research utilized two supervised machine learning algorithms known as logistic regression and random forest for the prediction of the stress level of the students. This decision was made due to the efficiency of the algorithms in classification problems as well as in generating reliable predictions.

A. Logistic Regression (LR)

Logistic Regression is a supervised machine learning algorithm which is popularly used for classification purposes. In this research study, it has been applied to classify the stress level of the students based on various parameters like Sleep Quality, Headache, Academic Performance, Study Load, and Extra-Curricular Activities. This algorithm learns from the relationship of the input parameters and the output target variable (stress level) and predicts the probability of the student falling into a particular stress category. The implementation of logistic regression algorithm is fairly easy and gives easily interpretable results.

B. Random Forest (RF)

The Random Forest is a supervised machine learning algorithm that works through ensemble learning technique. In other words, Random Forest constructs several decision trees, each with different sub-sets of training data and the output of the prediction results is derived through combining the outputs of all trees together. Unlike decision tree, which predicts using only one decision tree, Random Forest takes majority voting among all trees and thereby increases prediction accuracy.

This study applies Random Forest for predicting the stress level of students taking into consideration features like Sleep Quality, Headache, Academic Performance, Study Load, and Extracurricular Activities. With the capability of dealing with complexity among features and producing more accurate and reliable results, Random Forest outperforms the other algorithms and hence becomes the proposed model in this study.

8. System Flow

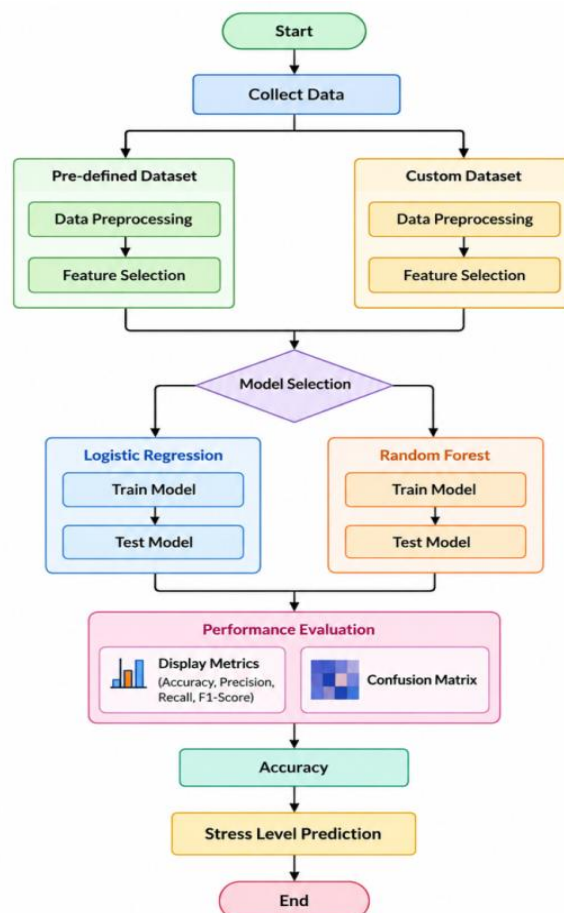


Fig 8.1: System Flow

Figure 8.1 shows the complete process followed to develop and evaluate the Student Stress Level Prediction system. In this process, first, the student stress dataset is collected and stored in CSV file format. Next, the collected dataset goes through the stage of data preprocessing in order to handle the issues like missing values, duplicates and inconsistent values if any. Following this, feature selection is done in order to find out relevant features, which include Sleep Quality, Headache, Academic Performance, Study Load, Extracurricular Activities, etc.

9. EXPERIMENTAL RESULTS

In this research, the dataset was formed in CSV form and used for prediction of stress level of students using machine learning methods. In this dataset, numerical data was provided which was concerned with different student stress parameters including academic performance, study load, sleep condition, headache, and extracurricular activities. Prior to building the model, dataset was preprocessed appropriately so that missing values could be eliminated and accuracy of predictions can be enhanced for classification of stress levels of students, two machine learning algorithms named Logistic Regression and Random Forest were used. Both the models were used for classification of stress levels of students and finding different patterns in the dataset. Compared to simple binary classification, classification of different levels of stress provides better and more precise idea about psychological stress of students.

Model	Accuracy	F1-Score	Precision	Recall
Random Forest	0.932692	0.932303	0.935948	0.932692
Logistic Regression	0.375000	0.339866	0.334942	0.375000

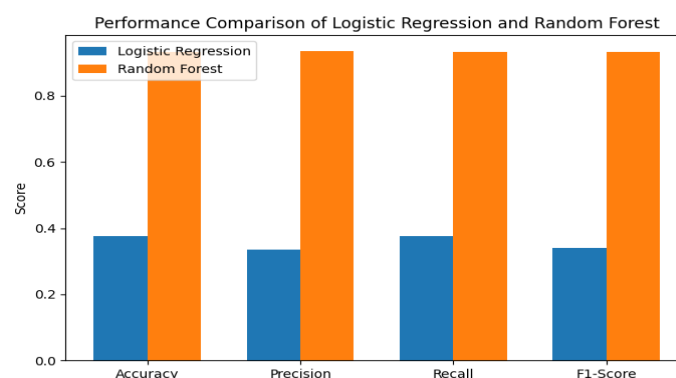


Fig: 9.1 of Performance comparison of Logistic Regression and Random Forest

Confusion Matrix

The confusion matrix was employed for evaluating actual versus predicted stress levels. In this dataset, we made use of the confusion matrices. Diagonal elements denote correctly classified instances, whereas non-diagonal elements reflect incorrect classifications.

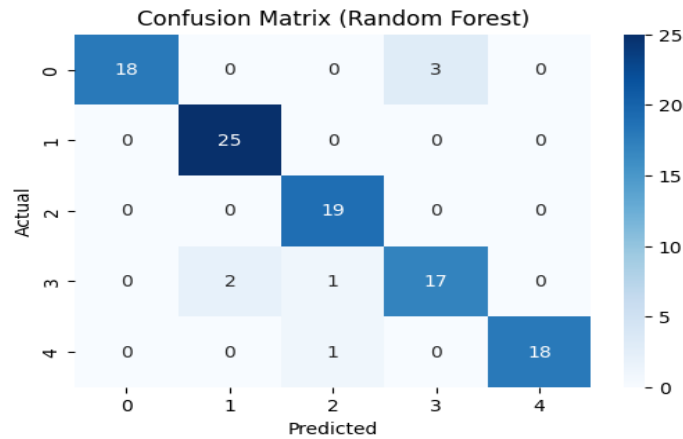


Fig 9.2 Confusion Matrix of Random Forest

Feature Correlation Heatmap

The following correlation matrix has been created to determine the correlation between each feature and the target variable. The heat map displays both positive and negative correlations among the variables. for instance

- Study Load correlates positively with the Stress Level
- Sleep Quality correlates negatively with the Stress Level.
- Headache correlates positively with the Stress Level.

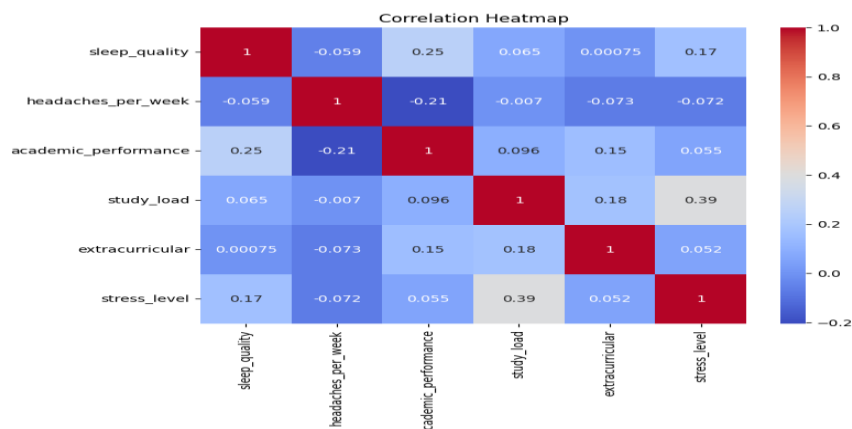


Fig:9.3 of Correlation Heatmap

Stress Level Distribution

The distribution of the target variable, Stress Level, was studied to see the number of students in different categories of stress. The data set has different stress levels from low stress to high stress. A bar chart was plotted

For the distribution of students in different stress categories.



Figure 9.4 of class distribution of stress level

Feature Importance

Analysis of Feature Importance was done by Random Forest.

Following were the top five features affecting students' stress:

- Study Load
- Sleep Quality
- Academic Performance
- Headache
- Extracurricular Activities

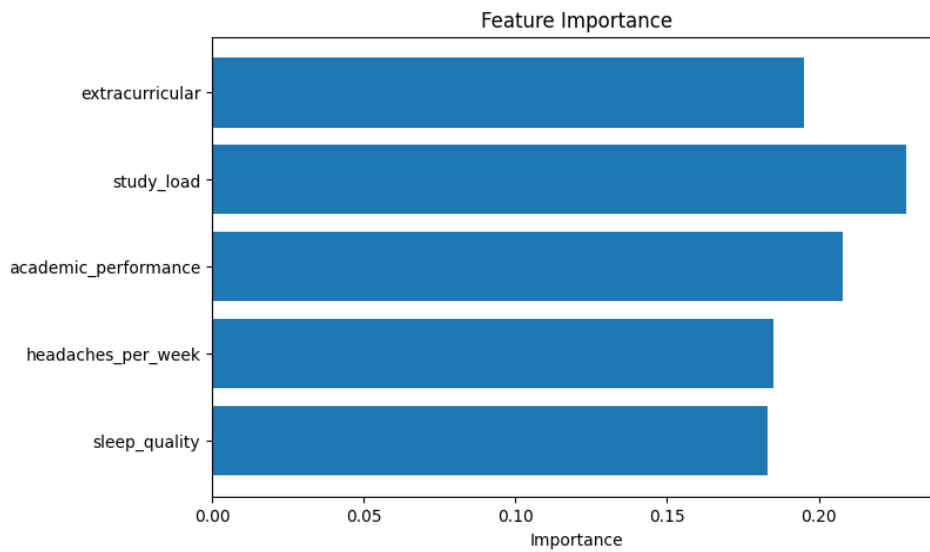


Figure 9.5 of feature importance

10. Conclusion

The study aimed at predicting the level of stress in students through the use of machine learning methods. A dataset with the variables of Sleep Quality, Headache, Academic Performance, Study Load, and Extra-curricular Activities were used in the experiment. Preprocessing of the dataset as well as statistical analysis of the variables was done before building the models.

Two supervised machine learning approaches, lr and rf, were implemented in order to predict the level of stress in the students. accuracy, precision, recall, and f1-score were used to calculate the performance of the two models. It was observed that both models were able to perform classification of student stress levels. However, Random Forest had more accurate predictions than Logistic Regression.

Based on the feature importance analysis, Study Load, Sleep Quality, and Academic Performance are some of the most important factors contributing to the stress level of students. As it was seen from the results, machine learning algorithms are able to efficiently evaluate various factors related to students' life and predict their stress level accurately.

To summarize, the Student Stress Level Prediction system suggested in this paper may help universities or colleges to detect the students with high levels of stress in time, which will enable them to take proper measures that might positively affect students' academic success and well-being.

In addition, based on the obtained results, the Random Forest model is an adequate tool for predicting student stress level due to its high accuracy and ability to handle complex relationships between variables.

11. Future Scope

The prediction of student stress was carried out by means of Logistic Regression and Random Forest. The results of the research are encouraging, yet the improvement of models is still possible.

In the future, it would be useful to obtain data from different educational establishments, where the model will have the possibility to get data on the students with different backgrounds and make predictions. Such aspects as social networks usage, family environment, financial situation, physical activity, and psychological state could be included since they may greatly affect the level of student's stress.

It would be appropriate to use more complicated machine learning and deep learning algorithms in order to improve the accuracy of predictions. Such algorithms as XGBoost, SVM, and Neural Networks may perform much better with the bigger dataset.

12. References

- [1] L. Rahunathan et al., "Machine Learning Model for Student Stress Level Prediction," in *4th OPJU International Technology Conference (OTCON)*, IEEE, 2025 [10.1109/OTCON65728.2025.11070804](https://doi.org/10.1109/OTCON65728.2025.11070804)
- [2] S.Dharaneedharan , "Analysis of Students Stress Level Using Machine Learning Algorithms," in *5th International Conferences On Expert Cloud and Applications(ICOECA)*,IEEE,2025 [10.1109/ICOECA66273.2025.00160](https://doi.org/10.1109/ICOECA66273.2025.00160)
- [3] G. Mahalakshmi and G. Sujatha, "Predictions of College Students' Mental Stress using Machine Learning Algorithms," in *7th International Conference on Intelligent Computing and Control Systems (ICICCS)*, IEEE, 2023 [10.1109/ICICCS56967.2023.10142672](https://doi.org/10.1109/ICICCS56967.2023.10142672).
- [4] Disha Sharma, Dr. Sandeep Singh Kang, "Stress Prediction of Students Using Machine Learning,"*International Journal of Mechanical and Production Engineering Research and Development(IJMPERD)*,2020 10(3).
- [5] Roocco de Filippis and Abdullah AI Foysal,"Comprehensive Analysis of Stress Factors Affecting Students: A Machine Learning Approach,"*Discover Artificial Intelligence*,2024.
- [6] Dr. Anusuya Biswas,"Modelling an innovative Machine Learning for Student Stress Forecasting,"*Global Perspectives in Management*,2024.