

Edu Mate AI: Smart Academic Assistant

Srushti Sanjay Sathe¹, Prajakta Sunil Shinde², Komal Pravin Waghmare³, Samrudhi Sharad Zadbuke⁴, Prof. Shalu Saraswat⁵

¹ PDEA's College of Engineering, Manjari, Pune, Maharashtra, India

² PDEA's College of Engineering, Manjari, Pune, Maharashtra, India

³ PDEA's College of Engineering, Manjari, Pune, Maharashtra, India

⁴ PDEA's College of Engineering, Manjari, Pune, Maharashtra, India

⁵ Professor, Department of Information Technology, PDEA's College of Engineering, Manjari, Pune, Maharashtra, India

Abstract - India has a large student population, making it difficult to monitor and predict their performance. Each educational institute has its own criteria for evaluating students, and there is no standard way of analyzing students' performance. Moreover, existing systems do not take into account significant external factors.

The proposed system, EduMate AI, aims to analyze and predict students' performance based on academic as well as non-academic factors, including geographical location, parental education, health conditions, and stress levels. Machine learning algorithms, Random Forest and XGBoost, are used to perform performance prediction and stress analysis.

The experimental results show that the inclusion of academic as well as non-academic factors enhances the accuracy of performance prediction, thus emphasizing the importance of data-driven approaches to improve academic performance monitoring systems.

Keywords: Artificial Intelligence (AI), Machine Learning (ML), Student Performance Prediction, XGBoost, Stress Detection.

1. INTRODUCTION

Education is a continuous and dynamic process, and identifying the academic strengths and weaknesses of students at the earliest is crucial to improving their performance. In higher education institutions, improving the quality of education is closely linked to the effective monitoring, analysis, and prediction of student performance.

The performance of students is affected not only by internal and external examination results but also by various personal and environmental factors. Factors like geographical location, education level of parents, family background, health, and stress levels are crucial in determining student performance. However, most existing systems are limited to academic performance and overlook such significant external factors.

The project aims to develop an AI-powered academic assistant, EduMate AI, to assist in academic management and student well-being. The proposed system will utilize facial recognition technology to enable efficient and secured login and Optical Character Recognition (OCR) to extract marks and attendance (to be implemented in the future). For prediction analysis, Random Forest and Extreme Gradient Boosting (XGBoost) are employed to increase the accuracy of predicting student performance and stress levels due to their efficiency in handling structured data. For the purpose of experimental analysis, data was collected from around 700 students from the Information Technology department. The data set consists of both academic and non-academic data, which makes it possible to analyze the performance and stress of students in a comprehensive manner.

The rest of this paper is divided into the following sections: Section II is dedicated to literature review, while Section III covers methodology, and Section V covers the conclusion with future scope.

1.1 Objective

The objectives of this project are to design an intelligent system that is capable of predicting student performance with high accuracy based on both academic and stress-related data.

The objectives of this project are:

1. To collect data related to academic and stress parameters of students.
2. To pre-process and clean up the data for further machine learning.
3. To design Random Forest and XGBoost models for performance prediction.
4. To compare models and analyze their performance.

The overall aim of this project is to design a system that is useful in improving academic monitoring and student stress in a reliable manner.

The proposed system aims to bridge the gap between performance analysis and stress detection by incorporating both in a single intelligent system.

2. LITERATURE SURVEY

A comprehensive literature survey of recent literature reveals various approaches for monitoring student performance and stress, mainly based on machine learning approaches. In previous studies, algorithms such as Random Forest and XGBoost were implemented to classify student performance. Even though good accuracy was obtained, manual input was required, and no automated process was implemented.

In another study based on stress prediction, research work was conducted utilizing smartphone sensor data along with the I-HOPE framework to estimate mental health. Even though good accuracy was obtained in the study, only behavioral patterns were considered. Moreover, academic performance variables were not included. In another study, mental stress was considered, but only survey data was used. Even though good accuracy was obtained in identifying stress levels, no relationship was established between stress levels and academic performance.

Further research was conducted on detecting mental stress based on questionnaire-based data. Even though good accuracy was obtained in the study, valuable insights were obtained about psychological conditions.

However, no relationship was established between academic performance and stress. This literature survey reveals that two separate approaches were implemented: one for stress prediction and another for academic performance analysis.

Other research, such as in [5], was conducted to predict academic performance based on demographic and external variables using Logistic Regression methods. This research emphasized the significance of non-academic factors in determining student performance. Nevertheless, this method was limited to general performance classification, without taking into account any subject-specific or psychological analyses.

In another research, various supervised machine learning algorithms were tested using data from two years of academic performance in the University of Basra. As per this research, Logistic Regression was found to have an accuracy of 68.7% for passed students and 88.8% for failed students, as mentioned in [6].

In conclusion, from the literature review, it is evident that considerable research has been carried out in predicting student performance and stress detection using various machine learning algorithms.

Nevertheless, most of this research is limited to individual analyses of both aspects, indicating a need for a comprehensive system that considers both aspects simultaneously.

3. METHODOLOGY

The proposed system is based on a systematic approach to measure student performance and stress level by applying various machine learning algorithms. The entire framework consists of data collection, preprocessing, feature selection, model development, and performance comparison. Both academic and non-academic data are considered to obtain precise and accurate predictions.

3.1 Data Collection

For experimental purposes, data was collected from around 700 students of the Information Technology department. Both academic and non-academic data are considered in this study, and both of them are included in the dataset.



Name	Age	Roll No	Gender	Passed Year	Current Year	10th Percentage	12th Percentage	Gap Years	Residence	Do you use your phone late at night when you have early classes	I feel my academic workload is overwhelming	I feel emotionally drained by my academic responsibilities	I feel lonely even when surrounded by classmates	I feel physically tired even when I haven't done much physical work	I feel supported
Ananya Das	18	1237	Male	2025	BE	85.0	78	0.0	Hostel	No	Neutral	Strongly Agree	Agree	Strongly Disagree	
Ashika Joshi	21	35		2025	BE	85.0	77	0.0	Hostel	Yes	Disagree	Strongly Disagree	Agree	Neutral	
Aruna	21	33	Female	2020	BE	80.0	85	0.0	Hostel	No	Agree	Strongly Disagree	Strongly Disagree	Disagree	
Aruna	21	33	Female	2025	BE	80.0	85	0.0	Hostel	No	Strongly Disagree	Neutral	Neutral	Strongly Disagree	
Ashik Zaidi	18	21	Male	2027	BE	80.0	83	0.0	Day Scholar	Maybe	Neutral	Agree	Disagree	Strongly Disagree	

FIGURE I: SAMPLE DATASET

3.2 Data Preprocessing and Feature Selection

Once the data has been collected, preprocessing is done to take care of missing values and make the data suitable for analysis. The appropriate operations such as data cleaning, normalization, and encoding are applied to the data. Feature selection is

done using correlation analysis to determine the most significant features that impact student performance. Data preprocessing is used for training the model and prediction.

3.3 Data Visualization

Data visualization is used for analyzing the relationship between various attributes in the dataset. Graphical representations are used to determine the dependency between academic and non-academic factors. Categorical data is converted into numerical form for analysis. Figures below represent examples of data visualization used in this study.



FIGURE II: STRESS LEVEL DISTRIBUTION

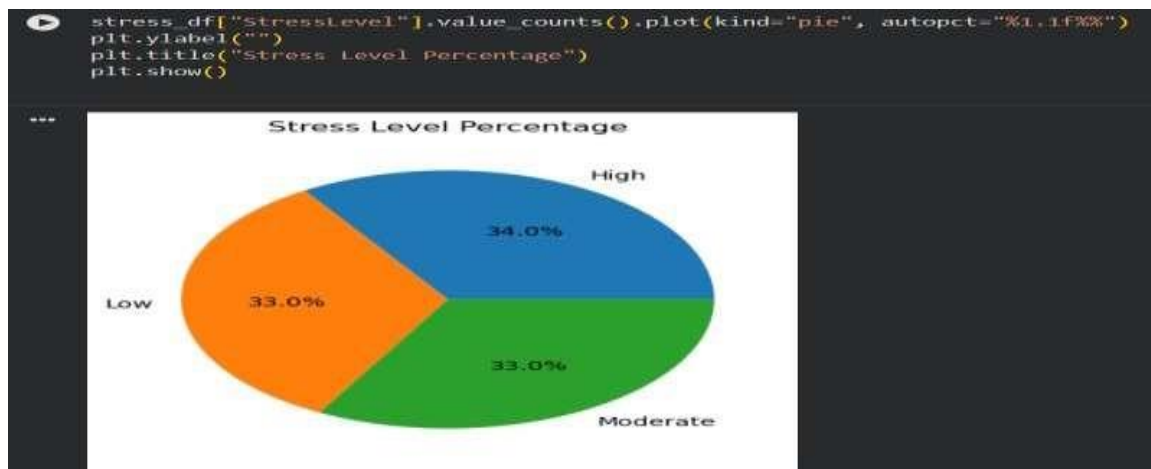


FIGURE III: STRESS LEVEL PERCENTAGE

3.4 Classification and Prediction

Two supervised machine learning algorithms, Random Forest and Extreme Gradient Boosting (XGBoost), are implemented for classification and prediction.

a) Random Forest Algorithm

The Random Forest algorithm is a supervised ensemble machine learning algorithm. It is mainly used for classification and regression problems. It creates multiple decision trees and combines their results to make predictions.

For this research, the Random Forest model has been implemented to classify students according to their performance levels.

The performance levels include:

- High
- Medium
- Low

The use of this algorithm has reduced the problem of overfitting and improved the accuracy of the model. It has achieved an accuracy level of 98%.



FIGURE IV: ACCURACY OF RANDOM FOREST MODEL

b)XG Boost Algorithm

XG Boost, also known as Extreme Gradient Boosting, is another supervised ensemble machine learning algorithm. It is considered one of the best algorithms for building decision trees. It creates decision trees sequentially. Each decision tree in XGBoost tries to correct the errors made by the previous decision trees.

For this research, the XGBoost algorithm has been implemented to classify student performance and predict stress. This algorithm has improved the accuracy of the model by minimizing the loss function. It has achieved an accuracy level of 94%.

```

learning_rate=0.07,
subsample=0.9,
colsample_bytree=0.9,
random_state=42,
use_label_encoder=False,
eval_metric='mlogloss'
)

model.fit(X_train, y_train)

y_pred = model.predict(X_test)

print("Test Accuracy:", accuracy_score(y_test, y_pred))
print("\nClassification Report:\n")
print(classification_report(y_test, y_pred, target_names=fe.classes_))

*** /usr/local/lib/python3.12/dist-packages/xgboost/training.py:199: UserWarning: [14:19:04] WARNING: /workspace/src/learner.cc:790:
Parameters: { "use_label_encoder" } are not used.

bst.update(dtrain, iteration=1, fobj=obj)
Test accuracy: 0.94193548130709677

Classification Report:
              precision    recall  f1-score   support

   High         0.94         0.94         0.94         53
   Low          0.98         0.94         0.96         51
  Moderate     0.91         0.94         0.92         51

 accuracy         0.94         0.94         0.94         155
  macro avg       0.94         0.94         0.94         155
 weighted avg     0.94         0.94         0.94         155
    
```

FIGURE V: ACCURACY OF XGBOOST MODEL

4. RESULTS AND ANALYSIS

The performance of the proposed models has been evaluated using various parameters such as accuracy, recall, and F1-score.

The comparison of the Random Forest and XGBoost models is as follows in

Table I. TABLE I: COMPARISON BETWEEN RANDOM FOREST AND XGBOOST MODELS

Criteria	RF	XGB	Conclusion
Accuracy	98%	94%	Random Forest slightly higher
Generalization	Moderate	High	XGBoost performs better
Stress Recall	90%	96%	XGBoost performs better
F1-Score	82-85%	88%	XGBoost performs better
Feature Importance	Moderate	Clear	XGBoost provides clearer insights
Pattern Handling	Good	Very Good	XGBoost handles complex patterns better

As per the results in Table I, the accuracy of the Random Forest model is higher at 98%, whereas the accuracy of the XGBoost model is 94%. However, the performance of the XGBoost model is higher in terms of recall and F1-score compared to the Random Forest model. This indicates that the XGBoost model has a higher potential for critical situations such as high stress levels. Moreover, the generalization ability of the XGBoost model is higher compared to the Random Forest model in handling complex situations. Therefore, even though the accuracy of the XGBoost model is slightly lower compared to the Random Forest model, the XGBoost model is more appropriate for real-time implementation. To further evaluate the classification performance of the models, confusion matrix analysis is performed.

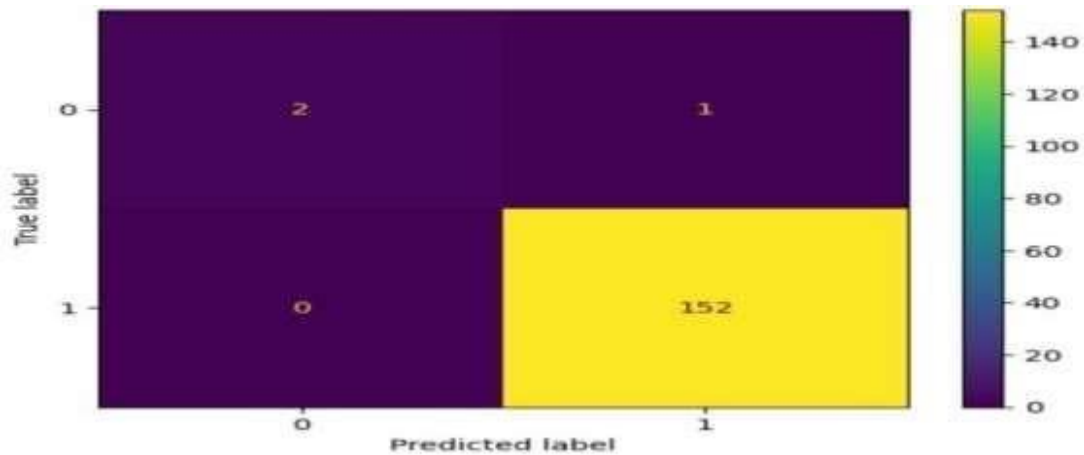


FIGURE VI: CONFUSION MATRIX FOR RANDOM FOREST MODEL FOR STRESS PREDICTION

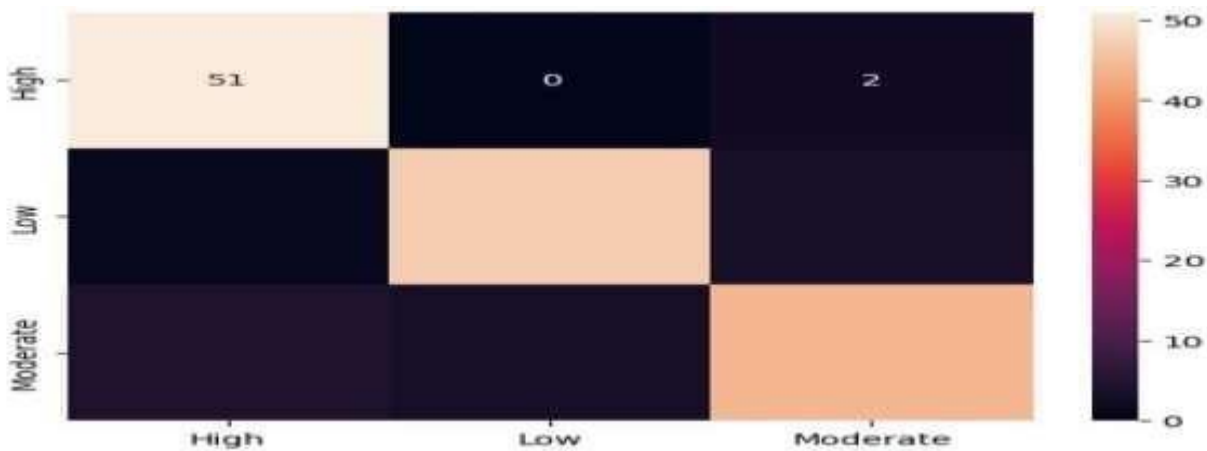


FIGURE VII: CONFUSION MATRIX FOR XGBOOST MODEL FOR STRESS PREDICTION

The confusion matrices of the two models are shown in Fig. VI and Fig. VII.

The confusion matrix of the Random Forest model indicates correct classification, but there are also misclassifications in the identification of at-risk students. However, the XGBoost model indicates better performance in classifying the students, with high true positive rates and fewer misclassifications.

This implies that XGBoost is better at detecting students at risk and dealing with complex data patterns.

5. FUTURE SCOPE

The future scope of the project includes the improvement of the intelligence and automation level of the system.

The XGBoost model can be trained with larger real-time data for better prediction accuracy and performance of the system. Based on the predicted academic results and stress levels, the system can be extended to provide a study plan for the student individually to improve their performance in the subjects in which they score low marks.

Moreover, the Optical Character Recognition technique can be implemented and integrated into the system for automatic extraction of marks and attendance from the uploaded document.

The system can be extended with additional features such as dashboards and the ability to predict the stress levels of the

student for a better academic support system.

Moreover, the system can be integrated with Learning Management Systems (LMS) for automatic synchronizing of the academic data for a complete academic support system.

Advanced predictive models can be integrated into the system for prediction of academic performance for a longer academic period and for the student individually for better academic support. These enhancements will transform EduMate AI into a comprehensive intelligent academic ecosystem.

6. CONCLUSIONS

In this study, the researchers employed various machine learning methods to analyze and predict the performance of students by utilizing structured data related to academics. The data was preprocessed and visualized, and two ensemble learning methods, Random Forest and XGBoost, were implemented for classification purposes.

The experimental results revealed that Random Forest had higher accuracy at 98%, while XGBoost had an accuracy of 94%. However, it was also revealed that XGBoost had significant advantages in terms of generalization and scalability, as well as the ability to handle complex data. Additionally, the model reduces the chances of overfitting.

Hence, even though the accuracy of the XGBoost model was slightly lower, it can be considered to be more appropriate for real-world applications due to its adaptability and consistency in all situations.

In general, the study revealed the effectiveness of ensemble learning methods in predicting student performance and also emphasized the importance of incorporating various advanced machine learning models in data analysis systems.

REFERENCES

- [1] Gupta, S., & Agarwal, J. (2022, October). Machinelearning approaches for student performance prediction. Paper presented at the 2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Amity University, Noida, India
- [2] N. T. M. Sagala, S. D. Permai, A. A. S. Gunawan, R. O. Barus, and C. Meriko, "Predicting Computer Science Students' Performance using Logistic Regression," Proc. 5th Int. Seminar on Research of Information Technology and Intelligent Systems (ISRITI), 2022
- [3] Anand Tripathi dept. of Artificial Intelligence and Data Science Datta Meghe Institute Of Higher Education and Research. Paper presented at the 2025 ,Dept. of Artificial Intelligence and Data Science Datta Meghe Institute Of Higher Education and Research.
- [4] Predicting and Understanding College Student Mental Health with Interpretable Machine Learning (2025) Citation: Meghna Roy Chowdhury et al., "Predicting and Understanding College Student Mental Health with Interpretable Machine Learning," CHASE '25, ACM/IEEE International Conference, 2025.
- [5] Machine Learning Algorithms for Detecting Mental Stress in College Students (2024) Citation: Ashutosh Singh et al., "Machine Learning Algorithms for Detecting Mental Stress in College Students," 2024.
- [6] Student Performance Prediction Approach Based on Educational Data Mining (2023, IEEE Access) - Ziling Chen, Gang Cen, Ying Wei, Zifei Li
- [7] Y. Wang, "Artificial Intelligence in Student Management Systems to Enhance Academic Performance Monitoring and Intervention," *Scientific Reports*, vol. 15, p. 35122, 2025.
- [8] H. Nellore, "Data Analytics and Machine Learning Approaches for Predicting Academic Stress and Enhancing Student Wellbeing," *International Journal of Advance Research and Innovation*, vol. 13, no. 4, pp. 23–29, 2025.