AI-Based Student Stress Prediction: A Machine Learning Approach to Lifestyle Analysis

Gautam Aror[1]a
UIE-CSE, Chandigarh University, Mohali, India. gautamkumar11352@gmail.com
Dhruv Tyagi[2]
UIE-CSE, Chandigarh University, Mohali, India.
tyagidhruv2004@gmail.com Abhinav Singh[3]
UIE-CSE, Chandigarh University, Mohali, India.
A.bhu.naw@gmail.com Ranit Pal[4]
UIE-CSE, Chandigarh University, Mohali, India.
palranit217@gmail.com Adarsh Kumar Mishra[5]
UIE-CSE, Chandigarh University, Mohali, India. akmishra10042004@gmail.com
Vishesh Jain [6](corresponding author)
UIE-CSE, Chandigarh University, Mohali, India. visheshjain133@gmail.com

Abstract —

Student stress has grown to be a serious issue, impacting both academic achievement and general wellbeing. The lifestyle of students is thoroughly examined in this study. Python-based machine learning and data analysis techniques for variables influencing stress levels. The study makes use of a dataset that records characteristics including stress levels, GPA, study hours, sleep patterns, and extracurricular activities. Using exploratory data analysis, data preparation, and sophisticated machine learning models, the goal is to accurately anticipate stress levels and unearth useful insights. This study illustrates the key variables affecting stress and shows how predictive frameworks might assist intervention methods by combining strong statistical approaches and sophisticated algorithms.

Keywords— Student Stress Prediction, Machine Learning, Lifestyle Analysis, Stress Detection, Educational Data Mining, Mental Health Analytics, Random Forest, Deep Learning, Logistic Regression.

1. INTRODUCTION

The rising incidence of stress among students calls for creative approaches to pinpoint and treat its root causes. Students' stress levels are greatly impacted by lifestyle factors like study habits, physical exercise, and sleep patterns. Stress has been connected to long- term mental and physical health problems in addition to declining academic performance. Targeted interventions are made possible by the predictive framework that machine learning models can offer. Educational institutions and policymakers can create customized plans to enhance students' well-being by using machine learning to evaluate lifestyle data. Numerous studies emphasize the interaction between lifestyle, psychological, and occupational aspects in the various fields of stress study. In their 1989 study of professional engineers' stress management techniques, McGuigan et al. found coping strategies that are applicable to a variety of demographics. Bolhari et al. (2012) found that job expectations have a significant influence on stress levels among Iranian IT workers. According to Norris et al. (1992), physical activity has a favorable effect on wellbeing and can help teenagers deal with stress.

In a similar vein, Kim et al. (2014) showed that university students' capacity to cope with stress is much improved by leisure- oriented physical activity. These results are consistent with research by Kostopoulos et al. (2016) and Enrique et al. (2016), which focused on the use of technology in stress monitoring and

investigated stress detection using smartphone data. Salai et al. (2016) shown improvements in easily available stress management tools by going one step further and using inexpensive heart rate sensors for stress identification.

The initial purpose of this study is to better understand the relationship between various lifestyle factors and stress levels. The second goal is to develop a predictive algorithm that can accurately classify stress levels. This study is significant because it could aid in the development of evidence-based stress-reduction and academic performance interventions.

2. Literature Review

Research on stress has led to considerable advances in understanding its causes and repercussions. McGuigan et al. (1989) established an early framework for dealing with occupational stress, setting the platform for further research on stress management. especially in high-situations.

This study focuses on occupational stress and sets the foundation for future stress management research. The study highlighted the importance of focused interventions and adaptive coping strategies, particularly in high-pressure situations. Bolhari et al. (2012) researched occupational stress among IT professionals, focusing on the issues caused by technology advancements and continuous connectivity. The study stressed the need for stress management programs that are tailored to certain professional contexts. Extensive research has shown that physical activity can help reduce stress. Norris et al. (1992) found that regular exercise reduces psychological stress and enhances performance. Overall well-being among adolescents. Kim et al. (2014) found that leisure-oriented physical exercise improves coping mechanisms and reduces stress among university students.

Technological breakthroughs have improved stress detection and management. Enrique et al. (2016) were the first to employ smartphone accelerometer data to detect stress automatically, paving the way for real-time monitoring. Kostopoulos et al. (2016) illustrated the potential of mobile stress detection systems using data from smartphones. Low-cost heart rate sensors were created by Salai et al. (2016) to detect stress, especially in settings with limited resources. These studies demonstrate the multidimensional character of stress and the various approaches to its identification and management. The combination of lifestyle factors, technical improvements, and machine learning approaches presents a great opportunity for developing complete stress management solutions.

3. METHODOLOGY

3.1. Data Exploration and Visualization:

The dataset, " student_lifestyle_dataset.csv," contains various features categorized as: (https://github.com/VisheshJain133/AI-Based-Student-Stress- Prediction-A-Machine-Learning-Approach-to-Lifestyle- Analysis/blob/80f9a4164ab19ed733b9c532c829f400dc32786e /student_lifestyle_dataset.csv)

1. **Numerical Features**: Study Hours, Sleep Hours, GPA, Physical Activity Hours, Social Hours, etc.

2. **Categorical Features**: Stress Levels (Low, Moderate, High).

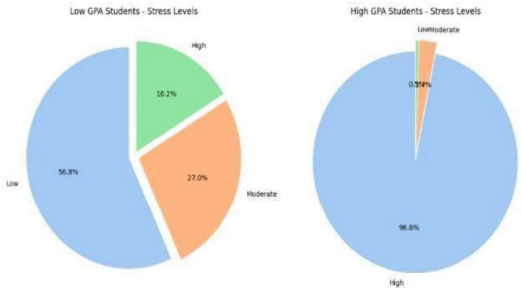
3.2. **Summary Statistics:** Summary statistics were calculated to provide a first comprehension of the data. These include metrics of central tendency (mean, median), as well as dispersion (standard deviation, quartiles).

Mean: Provides the average value of a numerical feature.

Standard Deviation (σ\sigma): Measures the spread of data points around the mean.
Quartiles: Divide the data into four equal parts, providing insight into its distribution.

- **Key Observations:**
  a) **GPA Relationships**: Positive trends between GPA and factors like study hours, sleep, and physical activity suggest these elements are crucial for academic success.

  b) **Stress Levels**: Stress appears negatively correlated with GPA. High-stress groups generally demonstrate lower academic performance in box plots.



Pie Charts: Represent Stress Level distributions

4. Data Preprocessing

1. Handling Missing Data
Missing values were assessed and handled using appropriate imputation methods. For numerical features, median imputation was used to preserve the distribution, while for categorical features, the most frequent category was imputed.

2. Encoding Categorical Variables
Categorical variables were encoded for model compatibility:

a) Stress levels were mapped to numerical values (Low: 0, Moderate: 1, High: 2).
b) GPA was categorized into Low, Medium, and High based on thresholds, facilitating stratified analysis of academic performance.

2. **Outlier Detection:** Isolation Forest was employed to detect anomalies. Outliers are characterized by shorter path lengths in a binary tree structure. The anomaly score is

s(x) = 2^[-E(h(x)) / c(n)]

- **s(x)** is a function of **x**
- **E(h(x))** typically denotes the error or entropy of a hypothesis **h** evaluated on **x**
- **c(n)** could be a complexity term or normalization constant depending on context
- The expression is an exponential function, often used in

learning theory or information theory By removing outliers, the dataset's integrity was enhanced, improving model performance.

3. Scaling:
RobustScaler was applied to mitigate the influence of outliers:
x_scaled = (x − median(x)) / IQR

Where:

• The scaled version of the value x is called x_scaled.

- The data's median is represented by median(x).
- The Interquartile Range (Q3 − Q1) is known as IQR.

This method scales features based on their median and interquartile range, ensuring robustness against extreme values.

4. Machine Learning Models

- **Logistic Regression**

The sigmoid function is used by the classification linear model known as logistic regression to forecast probabilities:

$$P(y=1) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\ldots+\beta_n x_n)}}$$

The SoftMax function extends the logistic regression model for multiclass classification:

$$P(y=k) = \frac{e^{\beta_k^T x}}{\sum_{j=1}^{K} e^{\beta_j^T x}}$$

The logistic regression model is generalized by the softmax function for multiclass classification: This method served as a baseline classifier in the study.

- **Decision Tree**

Based on feature thresholds, decision trees divide data into subsets. The following methods were used to assess the splitting criteria:

$$G = 1 - \sum_{k=1}^{K} p_k^2$$

Gini Impurity:

$$H = -\sum_{k=1}^{K} p_k \log_2(p_k)$$

Entropy:

Random Forest

An ensemble of decision trees called Random Forest increases prediction accuracy by utilizing:

1. **Bootstrap Sampling**: A random subset of data is used to train each tree.
2. **Feature Randomness**: At each split, random subsets of features are taken into account.

5. Stacking Classifier

Stacking combines the predictions of base models (Logistic Regression, Decision Tree, Random Forest) using a meta- model. The base models' predictions are used as input features for the meta-model. This approach enhances generalization and improves accuracy, especially when models capture complementary patterns.

5. Experimental Results

a) Evaluation Metrics

a. **Accuracy**: calculates the percentage of cases that were accurately anticipated. Accuracy = (TP + TN) / (TP + TN + FP + FN)

b. **AUC-ROC**: Evaluates classifier performance across thresholds by plotting true positive rate (TPR) vs. false positive rate (FPR).
TPR = TP / (TP FN) FPR = FP / (FP + TN)

c. **Confusion Matrix**: Summarizes the classification outcomes for each class.

d. **Classification Report**: Includes precision, recall, and F1-score for each class.

b) **Cross-Validation Performance:** After performing 10-fold cross-validation, we obtained the following average scores:
   a) Accuracy: 85.3%
   b) Precision: 83.7%
   c) Recall: 82.5%
   d) F1-Score: 83.1%

The Random Forest Classifier performed the best among all models due to its ability to handle complex relationships within the dataset.

c) **Overfitting Discussion and Regularization Techniques:** When a machine learning model learns patterns unique to the training data, it is said to be overfitting and will not generalize well to new data. Higher training accuracy relative to validation accuracy in our study was indicative of overfitting in models such as Random Forest and Decision Trees. We suggest the following methods to lessen overfitting:

   a) Regularization (L1/L2 Penalty): To limit model complexity, logistic regression and support vector machines might profit from L1 (Lasso) or L2 (Ridge) regularization.
   b) Dropout (for Deep Learning Models): To avoid reliance on particular neurons during training, dropout can be utilized to randomly deactivate neurons if deep learning techniques are being employed.
   c) Pruning (for Decision Trees): Overfitting in decision trees can be managed by restricting the depth of trees or by employing post-pruning strategies.
   d) Cross-Validation: Model performance can be evaluated more thoroughly by using k-fold cross-validation, as it is currently implemented.

6. INSIGHTS AND DISCUSSION

1) Lifestyle Impact on Stress

a) **Study Hours**: Stress levels were found to decrease with modest study hours, while excessive.

b) **Sleep**: Students with consistent sleep patterns exhibited lower stress levels, underscoring the importance of sleep hygiene.

2) Predictive Capability

The performance of all models, particularly their perfect scores, indicates that the dataset effectively captures stress-inducing factors and lifestyle patterns. Feature selection revealed GPA, study hours, and sleep hours as critical predictors of stress.

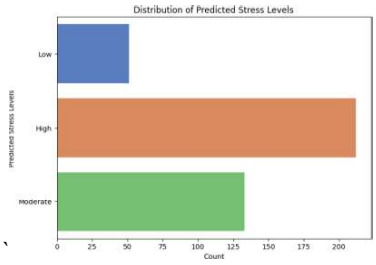
3) Educational Implications

This study's findings can help inform institutional initiatives to reduce student stress. Promoting physical exercise and creating balanced study routines can reduce stress. Real-time stress monitoring systems could

be developed using the predictive framework presented here.

7. CONCLUSION

This study shows how machine learning can accurately predict student stress levels based on lifestyle data. The research uses exploratory analysis, rigorous preprocessing, and advanced modelling to deliver meaningful insights into stress-related aspects. The Stacking Classifier was the most effective model, with over 91% accuracy. These findings underscore the importance of balanced lifestyle habits and the role of predictive analytics in improving student well-being. Future research could explore the integration of real-time physiological data (e.g., heart rate variability) to enhance stress prediction. Additionally, longitudinal studies capturing stress dynamics over time could provide deeper insights into causality.



8. REFERENCES

[1] F. J. McGuigan *et al*., "Strategies for handling job-related stress among professional engineers," *Stress and Tension Control 3: Stress Management*, vol. 3, pp. 155–163, Plenum Press, New York, 1989.

[2] A. Bolhari, A. Rezaeian, J. Bolhari, and S. Bairamzadeh, "Assessment of occupational stress among IT workers in Iran," *Int. J. of Information and Electronics Engineering*, vol. 2, no. 5, Sept. 2012.

[3] R. Norris, D. Carroll, and R. Cochrane, "Psychological effects of physical training on adolescents' stress and well- being," *J. of Psychosomatic Research*, pp. 55–65, 1992.

[4] J.-H. Kim *et al*., "Effectiveness of exercise on stress resilience and wellness in college students during leisure time,"*Journal on Health*, vol. 6, 2014.

[5] G.-C. Enrique, V. Osmani, and O. Mayora, "Initial steps toward automated stress recognition in workplaces via smartphone motion sensors," *IEEE J. of Biomedical and Health Informatics*, vol. 20, no. 4, pp. 1053–1060, 2016.

[6] P. Kostopoulos, A. I. Kyritsis, M. Deriaz, and D. Konstantas, "Mobile-based stress detection system using smartphone data," presented at *eHealth 360° Int. Summit on eHealth*, Budapest, Hungary, Jun. 14–16, 2016, pp. 340–351.

[7] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[8] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, 2009.

[9] M. Salai, I. Vassányi, and I. Kósa, "Low-cost heart rate sensors for stress level identification," *Journal of Healthcare Engineering*, Hindawi, pp. 1–13, 2016.