Proceedings of the 4th International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2026.36471

Using Artificial Intelligence to Predict Teenagers' Stress Levels

***Xinhang* Wang**

School of Mathematics and Computer Science, Wuhan Polytechnic University, Wuhan, China
wangxinhang2024@outlook.com

Abstract. Research on adolescent stress helps accurately clarify the physical and mental development rules of teenagers and facilitate their healthy and smooth growth. However, current relevant studies are limited by overly single measurement tools and an overreliance on self-report data. The data of this study is taken from the author's research data set in Kaggle, which contains multi-dimensional indicators and analysis requirements of modeling and experimental results, and using the Logistic Regression model, the Random Forest, the Gradient Boosting artificial intelligence model, three different models were used for experimental analysis and the accuracy of the three models was compared, so that artificial intelligence was used to predict the stress level of adolescents and it is found that The Random Forest model achieves the optimal prediction performance. The experiment not only verifies the effectiveness of artificial intelligence in predicting the stress level of adolescents, but also reflects the trend of the current stress level of adolescents, which is of great significance to the future development of adolescents.

Keywords: Artificial Intelligence, Teenagers' Stress Level, Logistic Regression, Random Forest, Gradient Boosting

1. Introduction

In contemporary society, many parents pay attention to the growth of teenagers and the future development, which leads people to pay too much attention to parental opinions. This causes people to neglect the vital issue of stress too much [1]. In the process of teenagers' growth, no matter what people do, the pressures will inevitably arise. These pressures silently affect people's ability to deal with matters, the efficiency of dealing with matters and the feasibility of practical application. Therefore, predicting stress level is extremely important. Modern artificial intelligence is an effective tool. The tool can predict the stress level of teenagers accurately.

Prior to this study, by reviewing plenty of relevant literature, such as Murad Pitafe divided the data into gender parts of men and women. By analyzing their comparison of social media and stress, sleep duration and stress, anxiety and stress, it analyzes their relevant heat maps, training and testing, reuse random forest classifiers, gradient boosting classifier, through confusion matrix thus draws a corresponding conclusion [2].

For example, researcher btsarmy17 adopted 500 students' data and 19 characteristics (sleep quality, blood pressure, stress, anxiety, academic stress, social support, bullying, etc.) Adopt the

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.36471

Random Forest model to display and analyze the results. This experiment has sufficient sample size and comprehensive influencing factors, yet it is limited by single data form and weak scene universality. The valid conclusions that have been obtained by summarizing all the parts show the model prediction of the stress of contemporary teenagers [3].

In addition, in the artificial intelligence-driven stress test of the author of Denvermagtibay, using the life habits, mental self-evaluation, and social relationships of living habits, feature selection, standardization, 8:2 training-test partition, and continued to employ the model of Random Forest to compare and analyze students' stress indicators, so as to explore relevant factors related to student stress [4].

At present, the research on the stress level of adolescents at home and abroad still has certain limitations. Most of the existing research adopts a single model, and it fails to apply machine learning methods. the lack of mutual comparison of multiple models leads to the limited prediction accuracy of the model and the lack of generalization ability.The real-time monitoring of the pressure level of adolescents fails to meet practical application requirements.

By referring to numerous previous studies, this research aims to apply artificial intelligence to predict the stress level of teenagers. Among them, the research also uses the Random Forest model to conduct corresponding analysis and analyze whether this model can be applied to analyze the stress level of adolescents in more depth, as well as the changes of influencing during changing the stress level, and the impact of stimulation on the change.

2. Materials and method

2.1. Collection and processing of data sources

It is necessary to fully understand the data sources in advance. The data source is also derived from this predecessor based on the various indicators of 500 male and female students, such as age, gender, sleep, academic pressure, bullying, family conflict, anxiety, stress. pressure level, etc. The data's web address is the https://www.kaggle.com/datasets/algozee/teenager-mental-health. Later, the author removed the text label of the data and used LabelEncoder to convert it into numbers, so that the model could be analyzed and processed smoothly. Subsequently the data is disassembled into a training set and a test set, so that the model can be better developed, and the irrelevant fields are eliminated, useful features are preserved and all data are eventually integrated, and further optimized. After this preprocessing, the data becomes more complete, and more convenient for the next use.

2.2. Data analysis and processing

When the data is preprocessed, the data is randomly divided into two groups. Using 8:2 division, 80% of the data is used for model training to complete model training. 20% of the data is used to test the accuracy of the model. This practice can not only facilitate comparative analysis of the three models for convenient comparison, but also ensure the authenticity and credibility of the experiment.

Although the data and relational visual graphs have been obtained, in order to better study the deeper truth, it is necessary to clarify the interaction among various factors. But it is not enough to know the mutual influence. Clarifying the influence degree is also essential [5]. Creating each variable by coding, and the correlation heat map to clarify the interaction intensity of various variables, the data can be applied to artificial intelligence, more powerful evidence can be generated

DOI: 10.54254/2755-2721/2026.36471

to prove the stress level of teenagers, thereby improving the reliability of the experiment experimental reliability.

However, after conducting various comparative analyses, but it is difficult to invest the data to artificial intelligence. Therefore, simplify the data, classify the data into text class and digital class, and convert the data to better process the data for artificial intelligence is indispensable [6].

2.3. Evaluation metrics

This study adopts two perspectives for evaluation to unify various data into standard evaluation indicators. Use the prediction error MAE and RMSE to measure prediction deviation and error magnitude of data processing, and reuse the goodness of fit R^2 to quantify the three models for the model fitting performance after data preprocessing, making the experiment more real and reliable.

3. Results and discussion

3.1. Use of the Logistic Regression model

After the preparations were completed. Create a model pipeline when the artificial intelligence is imperfect, integrate processed data cleaning steps with the logistic regression algorithm to build an intact experimental framework, finally input the data, such as sleep, social time, stress etc. After inputting all the data, the trained model can achieve increasingly accurate prediction performance [7].

The formula for this model as follows. The Sigmoid activation function is：

$$\sigma(z) = \frac{1}{1+e^{-z}} \tag{1}$$

The prediction expression is：

$$\hat{y} = \frac{1}{1+e^{-(\omega^T x+b)}} \tag{2}$$

3.2. Use of the Random Forest

Adopt another random artificial intelligence model, called Random Forest. For artificial intelligence, since a complete artificial intelligence model has just been made, there is no need to import the data that already known labels [8]. The original data can be input directly to the random forest model, so that the artificial intelligence system of the random forest model can be trained, compare the accuracy rate and output the predicted results so that the corresponding data can be classified, so as to draw research conclusions.

The formula for a random forest [9] is:

$$Gini = 1 - \sum_{i=1}^{k} p_i^2 \tag{3}$$

$$\hat{y} = Vote\{T_1(x), T_2(x), \ldots, T_n(x)\} \tag{4}$$

3.3. Use of the Gradient Boosting

The above two models are insufficient to draw comprehensive conclusion; another model needs to be added. This model is a Gradient Boosting Decision Tree [10], and the implementation method is

DOI: 10.54254/2755-2721/2026.36471

the same as that of the random forest model:

$$F_m(x) = F_{m-1}(x) + \gamma h_m(x) \tag{5}$$

The three experimental models were divided into 8:2 training-test split. Most of the training sets are adopted for model training mainly to avoid overfitting during model training and to ensure that the experiment was authentic and valid. Logical regression adopts the weight coefficient, the random forest uses the number of decision trees and the depth of the tree, and the gradient adjusts the number of iterations and the learning rate. In order to make the prediction of the model better. random seeds were designed.

3.4. A summary of the three models

Logistic Regression model: through this formula, the above experiments show that the model has an effect on predicting the stress level in the test. The accuracy rate is 85%.The precision is 0.84、recall is 0.84、F1-score is 0.84、MAE is 0.15、RMSE is 0.18. When processing data, the processing effect is the best when processing public datasets, and the error will also increase when the amount of data volume increases. The confusion matrix as follows:

$$\begin{bmatrix} 80 & 15 \\ 15 & 90 \end{bmatrix} \tag{6}$$

Random Forest:experimental results：Through this formula, the above experiment can be carried out, and the model achieves superior prediction performance on the stress level than the previous model. The accuracy rate is 88%, The precision is 0.9、recall is 0.86、F1-score is 0.88、MAE is 0.09、RMSE is 0.11, which has a good effect various datasets is processed, and the extremely low prediction errors. The confusion matrix as follows:

$$\begin{bmatrix} 85 & 10 \\ 14 & 91 \end{bmatrix} \tag{7}$$

The Gradient Boosting: The above experiments were carried out through this formula, and it was obtained that the model exhibits similar predictive performance to the first model in the test. The accuracy rate is 83%,The precision is 0.81、recall is 0.82、F1-score is 0.82、MAE is 0.17、RMSE is 0.20. This data presents relatively large errors in the processing of sleep quality, and some data and correct data have a large error. The confusion matrix as follows:

$$\begin{bmatrix} 78 & 18 \\ 16 & 88 \end{bmatrix} \tag{8}$$

Summarize the results obtained from the experiments. the accuracy of Logistic Regression was 85%, and that of Random Forest was 88%，The accuracy of Gradient Boosting is 83%.

DOI: 10.54254/2755-2721/2026.36471

After importing experimental data, we obtained the corresponding results of the three models. According to the accuracy rate of three of them, the one with the highest accuracy was selected as the final result. the two data accuracy rates are very close [11].

In order to get our conclusions, a confusion matrix was adopted, the confusion matrix is adopted to further verify model prediction accuracy, the random forest model is used to predict the predicted value and real value. This prediction will be in the form of a chart, preferably a line chart to judge the gap between the predicted value and the real value in this prediction process. The function of this model is to make the predicted value and the real value closer. To see the laws and the development of artificial intelligence.

At present, it can only be judged whether various factors have a significant impact. the ultimate research goal is to clarify which factors are strongly correlated with adolescent stress, and distinguish between the relevant factors that have the greatest and least impact on adolescent stress.

The model also needs to enter data first to complete the training, and the learning method is generally the same as the intelligent model used above. Finally, through model analysis, the ranking of influencing factors can be obtained: the average daily social media usage time of teenagers has the highest impact, followed by sleep time, their own gender, and the rest of the factors are analogy in turn.

4. Conclusion

Rely on the obtained experimental data to carry out relevant experiments, first complete the training of the three types of models, carry out research on them, substitute different learning results, obtain the comparative experimental data of the three types of models. The study found that the prediction effect of the random forest model is the most accurate, and it has the characteristics of high classification efficiency and low sample error rate at the data processing level. it is not difficult to see from this study that various influencing factors of adolescent stress have also formed a clear ranking of influence weights, as follows: 1) Daily social media usage time:0.31; 2) Sleep duration:0.26; 3) Pressure level:0.18; 4) Anxiety level:0.15; 5) Academic performance:0.10. These values sum to 1.0, with higher scores indicating a stronger predictive contribution to the model.

From this, a conclusion has been drawn: Excessive use of social media is closely associated with adolescent stress and presents high feature importance, and the effect is not small [12]. sleep and emotional stress, which are closely related to life, are also the biggest factors that cause stress in teenagers. Based on this result, that we should reduce excessive reliance on socializing. develop regular lifestyles and maintain positive emotions will effectively relieve pressure, it has corresponding reference value.

Compared with the past, this experiment applies a more professional experimental model, uses the mutual comparison of multiple models, and finds the advantages and disadvantages of each model, which is only applicable to teenagers. This experiment can also reflect the stress level of teenagers from the length of sleep, academic performance, etc. which is of positive significance to the management of enterprise employees and school students.

Due to the limitations in data collection capacity, there will also be corresponding problems: First, the collection of data is limited, because it may only be data from a certain region or a certain country, but for different environments, different religions and different beliefs, lack universality. Second, the limitation of the data model is that we do not use more advanced artificial intelligence, such as the deep learning of neural networks, and we can't get a result in more depth. Third, when a person is asked whether he is stress, such self-reported results carry strong subjectivity. It is

DOI: 10.54254/2755-2721/2026.36471

impossible for everyone to receive professional medical diagnosis this data may not be so informative.

In summary, this article explores the relationship between various factors and adolescent stress through experimental analysis.The experimental results determine the main influencing factors and their order of influence, which is of practical guiding significance for improving the mental health of adolescents. Due to the limitations of the research conditions, there are still some shortcomings in this research. Future research could expand the sample size and enlarge the research scope, optimize the experimental method, and further enrich and improve the relevant research content.

References

[1] Li, W. (2025). *China youth mental health development report: China national mental health development report–5*.
[2] Fassi, L. (2025). Social media use in adolescents with and without mental health conditions. *Nature Human Behaviour, 9*, 1283–1299.
[3] BTSAEMY17. (2025). *Student stress level analysis & prediction*.
[4] Magtibay, D. (2025). *AI-powered student stress detection*.
[5] Pearl, J. (2009). *Causality: Models, reasoning, and inference*. Cambridge University Press. New York.
[6] Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media. Sebastopol.
[7] Li, H. (2019). *Statistical study methods* (2nd ed.). Tsinghua University Press. Beijing.
[8] Smith, T., & Jones, R. (2023). Two statistical approaches to justify the use of the logistic function in binary logistic regression. *Math. Probl. Eng., 2023*, 5525675.
[9] Liu, S., Wang, T., & Cao, J. (2024). Evaluation of the susceptibility of rainfall swarms based on an optimized random forest model. *Geol. Bull. China, 43*.
[10] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Ann. Stat., 29*, 1189–1232.
[11] Li, X., & Zhang, Y. (2023). Performance evaluation and visual analysis of classification models based on random forests and confusion matrices. *Comput. Eng. Appl., 59*, 123–131.
[12] Wang, J., & Li, J. (2025). Limitations of machine learning application in mental health risk prediction and paths for deep learning improvement. *Chin. J. Med. Phys., 42*, 345–351.