DOI: 10.23977/appep.2025.060123
ISSN 2523-5842 Vol. 6 Num. 1

Research on Children's Mental Health Assessment Model Based on Machine Learning

Jiao Tang[1], Qingkun Yu[2,*]

[1]School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, 114051, China
[2]College of Science, University of Science and Technology Liaoning, Anshan, 114051, China
[*]Corresponding author: yuqingkun@163.com

Keywords: Machine Learning, Smote Over-sampling, Principal Component Analysis, Children's Mental Health

Abstract: In the current context, where children's mental health problems are increasingly severe and traditional assessment tools are highly subjective, inefficient, and have poor predictive capabilities, this study focuses on improvement. Data from a specific website is collected and undergoes fine preprocessing. SMOTE is applied to increase the proportion of relevant data from 0.125 to 0.250, enhancing data balance. PCA is used for dimension reduction to identify emotional load and social context as key predictors. After screening multiple regression models, the GBDT model is selected. The R ^{2}values of the training and test sets are 0.999 and 0.995, respectively, with the MSE of the training set being 0.250 and the MAE being 0.374, and the MSE of the test set being 0.457 and the MAE being 0.233, showing low prediction errors. The innovation lies in integrating advanced techniques and algorithms to provide a more accurate and reliable assessment method for children's mental health, which is significant for promoting research in this field.

1. Introduction

In today's rapidly evolving social landscape, the issue of children's mental health has become alarmingly prominent. The escalating prevalence of problems such as depression, anxiety disorders, and behavioral anomalies has cast a long shadow over children's growth trajectories. It has undermined their academic pursuits, disrupted their social integration, and taken a toll on their physical and mental well-being. This has, in turn, sparked intense concern across society and urgently called for more refined assessment modalities. Historically, traditional methods of evaluating children's mental health have been predominantly subjective, highly inefficient, and woefully lacking in predictive acumen. Machine learning has presented itself as a glimmer of hope in the digital era. However, it has been beset by many challenges in real-world applications. Sample imbalance, high-dimensional data complexity, and intricate feature correlations' entanglement have all posed significant stumbling blocks.

This study aims to engineer a robust and effective children's mental health assessment model. Its significance is profound. Facilitating the early detection of mental health issues can trigger timely interventions and enable the continuous tracking of children's long-term development. This will indubitably foster children's healthy maturation and catalyze society's harmonious evolution. The novelty of this research resides in the seamless integration of a suite of cutting-edge data processing techniques and machine learning algorithms. Initially, an exhaustive collection of multi-dimensional data related to children's mental health is amassed from professional data repositories. This encompasses a wide gamut of factors, including but not limited to depression levels, anxiety manifestations, behavioral patterns, family dynamics, and peer interactions. Subsequently, a rigorous data preprocessing regimen is executed. Box-and-whisker plots are employed to meticulously detect and manage outliers, safeguarding the integrity of the data. The SMOTE oversampling technique is deployed to redress the sample imbalance conundrum, augmenting the relative abundance of pertinent data. Principal Component Analysis (PCA) is harnessed to distill the essential features, reducing the original five variables to the key dimensions of emotional load and social context. Finally, through an exhaustive comparison of regression models such as KNN and GBDT, the GBDT model is singled out as the preeminent choice. It offers a formidable scientific underpinning for children's mental health assessment and serves as a potent catalyst for the advancement of research and practice in this crucial domain.

Moreover, the research findings can inform the development of evidence-based mental health policies and programs tailored to children. Providing a more accurate and objective assessment tool can help mental health professionals allocate resources more efficiently and design targeted interventions. This, in turn, can lead to improved outcomes for children struggling with mental health issues and contribute to the overall well-being of future generations.

2. Data Sources and Preprocessing

This study is dedicated to the assessment and research of children's mental health, with data obtained from the website http://tjjmds.ai-learning.net.The website provides multidimensional data covering key information such as the Child Depression Index, Generalised Anxiety Disorder Index, Child Behaviour Scale Index, Family Relationships Index, and Peer Relationships Index, as well as background variables such as gender and age, to lay the groundwork for subsequent research.

Data preprocessing was performed immediately after data acquisition. For the indices of child depression, generalized anxiety disorder, child behavioral scales, family relationships, and peer relationships, their original scoring systems were converted to a new system with a score of 0, representing the least severe degree. On this basis, the scores of these indices were further aggregated to form mental health rubrics, thus providing a more targeted basis for analysis in subsequent studies. At the same time, gender and other variables are converted into 0-1 type numerical variables, and this treatment effectively simplifies the subsequent operation process and reduces unnecessary complexity.

However, during the data processing stage, this study found that there was a category imbalance problem that had a significant negative impact on the performance of traditional classification algorithms. To effectively address this challenge, the SMOTE [1] (Synthetic Minority Over-sampling Technique) oversampling technique was used in this study. Specifically, for non-numeric data, the One-Hot Encoding [2] method is applied to optimize the class distribution of the data by transforming each non-numeric feature column into multiple binary feature columns, each corresponding to one value of the original feature.

As shown in Figure 1, after the SMOTE oversampling process, the proportion of the sample data was increased from the original 0.125 to 0.250, which covers data related to children's mental health rating values. This adjustment has significantly enhanced the balance and representativeness of the data, providing more reliable data support for subsequent in-depth research and analysis and contributing to a more accurate assessment of children's mental health.

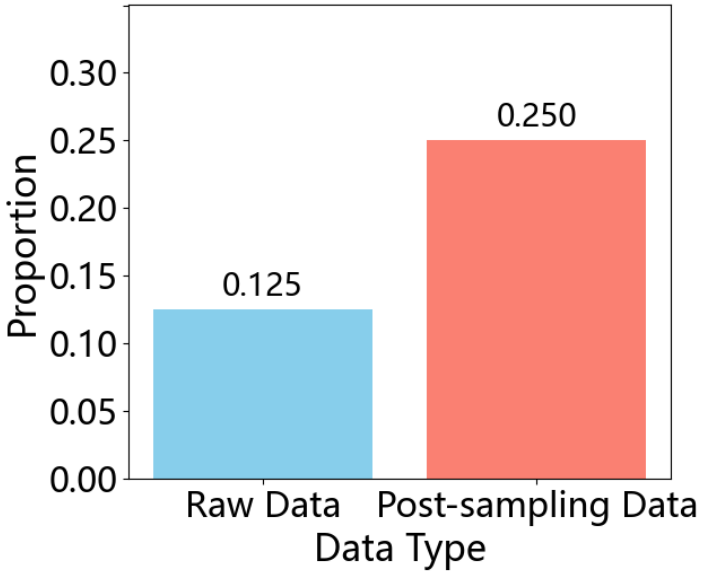

Figure 1. Ratio of SMOTE over-sampled data to original data

3. Principal Component Analysis and Data Dimensionality Reduction

Determining the suitability of data for principal component analysis is a critical step in the data exploration and feature engineering process. It is usually determined using Bartlett's test of sphericity. Table 1 below shows the results of this study regarding Bartlett's sphericity test.

Table 1. Bartlett's test of sphericity

KMO value	0.651
approximate chi-square value	1818.532
degree of freedom	10.000
p-value	0.000

As seen from Table 1, the KMO value of 0.651 is in the range of 0.6-0.7, which aligns with the requirements of the principal component analysis, indicating some correlation in the data. A p-value of 0 less than 0.05 indicates that the correlation matrix of the data is a non-unit matrix, proving that the data is suitable for principal component analysis. Its approximate chi-square value of 1818.532 and degree of freedom of 10 highlights that the data correlation is significant.

Based on these results, the five variables of depression, anxiety, problem behavior, family upbringing, and peer relationships can be downscaled using principal component analysis. This operation aims to dig deeper into the underlying structure of the data to develop scientifically sound indicators for judging mental health. Figure 2 below further demonstrates the information related to principal component analysis with the help of a gravel plot.

As can be seen from Figure 2, downscaling the five variables into two components is the best choice. This is because the first component has a large eigenroot of more than 2.0, the second component also has a relatively large eigenroot of about 1.0, and the eigenroot decreases significantly from the third component onwards. According to the rule of thumb that the eigenroot is greater than 1, the eigenroot of the first two components meets the requirements. The change in the slope of the curve tends to flatten out at the third component, which indicates that the first two components have been able to capture most of the variation information of the data, and adding more components will have a limited incremental increase in the explanation of the data variance, so the data are downscaled to 2 components using PCA [3]. Naming these two principal components as Emotional Load Indicator and Social Context Indicator, respectively, helps the subsequent machine learning model to select appropriate input features, thus improving model performance and efficiency.

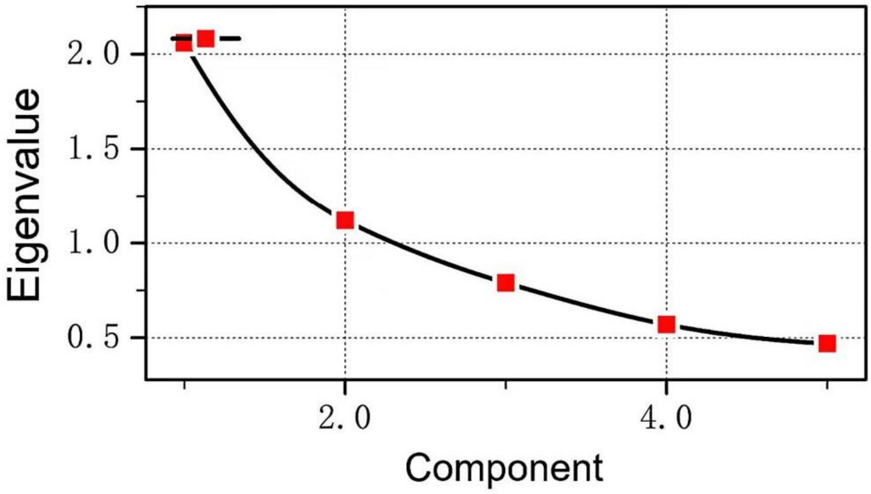

Figure 2. Gravel Plot

4. Regression model performance comparison and model selection

Before the screening regression model predicts CMHEV, it is scientifically classified into different degrees according to a specific total value range. This division integrates multidimensional factors covering internal emotions such as depression and anxiety, external manifestations such as problematic behaviors, and fathering and peer relationships to provide a comprehensive picture of children's mental health. The results of the correlation division are presented in a concise and intuitive table, and the range of the degree of judgment for each variable is detailed in Table 2.

Table 2. Range of levels of judgment for each variable

Type	Good	Moderate	Cause for Concern	Severe
CMHEV	0.00-48.75	48.76-97.50	97.51-146.25	146.26-195.00
Depression	0.00-13.50	13.60-27.00	27.10-40.50	40.60-54.00
Anxiety	0.00-5.25	5.26-10.50	10.51-15.75	15.76-21.00
Problem Behaviour	0.00-8.00	8.01-16.00	16.01-24.00	24.01-32.00
Fathering	0.00-22.00	22.01-44.00	44.01-66.00	66.01-88.00
Peer relationships	0.00-8.00	8.01-16.00	16.01-24.00	24.01-32.00

Next, appropriate models need to be screened to predict CMHEV through emotional load indicators and social context indicators and then to classify children's mental health in terms of degree according to the range of degrees of classification. This study compares the effect and performance of KNN [4] and GBDT [5] to select the best model.

First, look at the test data prediction graphs; the test data prediction graphs for the KNN and GBDT comparison are shown in Figure 3 and Figure 4 below, where blue represents the actual value and green represents the predicted value.

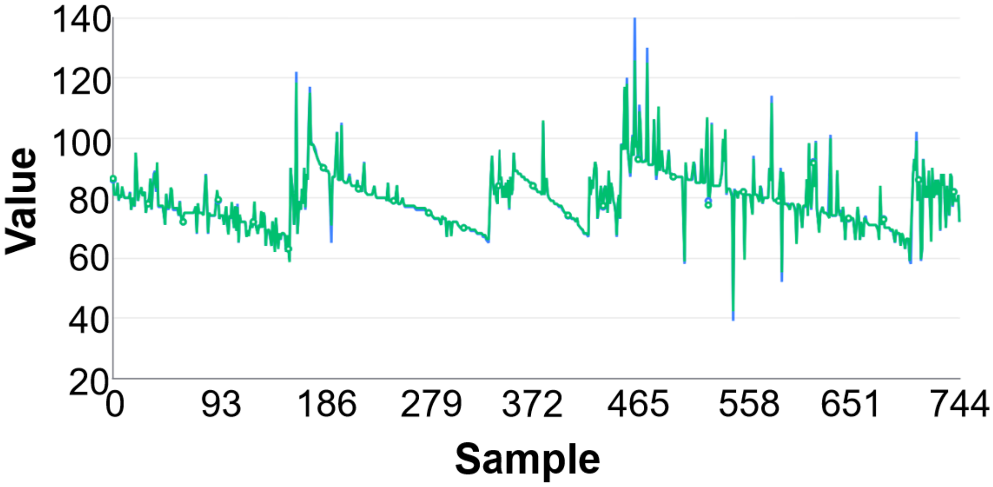

Figure 3. KNN prediction scenarios

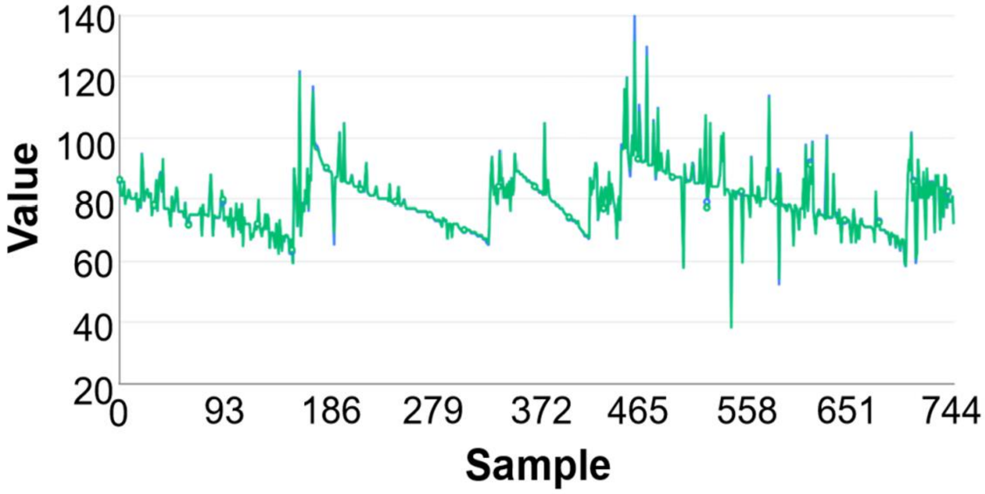

Figure 4. GBDT prediction scenarios

From the test data prediction plots in Figures 3 and 4, the green prediction line of KNN can roughly follow the fluctuation trend of the blue actual value line. Still, the deviation between the two is apparent in many sample points, especially in the fluctuating region. The prediction value fails to match the change of the actual value accurately. In contrast, the prediction line of GBDT has a higher degree of fit to the exact value line and is smore closely related to the direction of the actual value. In terms of fluctuation, the fluctuation of KNN's predicted value folds is frequent and significant, showing that it is over-sensitive to local data changes or there is an over-fitting phenomenon. GBDT's predicted value folds are relatively smooth, with minor fluctuations, reflecting the advantage of stability in processing data. Regarding prediction accuracy, the KNN model has a relatively large prediction error due to large fluctuations in the predicted value and many deviations from the actual value. In contrast, the expected value of the GBDT is closer to the actual value, and the overall prediction accuracy is higher.

Tables 3 and 4 below show the results of the KNN and GBDT models, taken together with the specific evaluation metrics data.

Table 3. KNN evaluation results

Data	MSE	RMSE	MAE	MAPE	R ²
Training set	2.483	1.576	0.874	1.21	0.992
Test set	0.705	0.839	0.277	0.341	0.992

Table 4. GBDT evaluation results

Data	MSE	RMSE	MAE	MAPE	R ²
Training set	0.25	0.5	0.374	0.543	0.999
Test set	0.457	0.676	0.233	0.312	0.995

As can be seen from Tables 3 and 4, in terms of the training set, the MSE (2.483), RMSE (1.576), MAE (0.874), and MAPE (1.21) of the KNN have significantly higher values compared to their GBDT counterparts, 0.25, 0.5, 0.374, and 0.543; at the same time, the coefficient of determination of the KNN is 0.992, which is lower than that of the GBDT of 0.999. This shows that GBDT has higher fitting and prediction accuracy on training data. Switching to the test set, the MSE (0.457), RMSE (0.676), MAE (0.233), and MAPE (0.312) of GBDT are lower than that of KNN and its coefficient of determination is 0.995, which is higher than that of KNN's 0.992. This indicates that GBDT also performs well on the test data, with more minor prediction errors, better fitting, and better generalization to predict new data better.

After careful consideration, GBDT has significant advantages. The prediction graph shows that its fit is higher, the fluctuation is smooth, and the prediction accuracy is substantial; the evaluation metrics show that all the key metrics are better than KNN on both the training and test sets. Therefore, GBDT can accurately grasp the data patterns and effectively respond to new data. It is a more reliable and appropriate model choice for subsequent work on children's mental health degree classification and intervention.

5. Conclusions

This research on the children's mental health assessment model represents a substantial effort considering the exacerbating problems in children's mental health, the limitations of traditional assessment means, and the data difficulties in machine learning applications. Extensive data related to children's mental health is painstakingly collected from a specific website and then undergoes a comprehensive and detailed preprocessing procedure. Through the application of SMOTE, the data balance is effectively enhanced by increasing the relevant data portion. Bartlett's sphericity test verifies the data's suitability for principal component analysis, and five variables are skillfully reduced to two crucial indicators, namely emotional load and social context. By systematically comparing regression models such as KNN and GBDT, the GBDT model demonstrates outstanding performance regarding goodness of fit, stability, and prediction accuracy, firmly establishing its feasibility and effectiveness in children's mental health assessment.

However, it is acknowledged that the study has certain limitations. The reliance on a single data source restricts the diversity of the sample and potentially impacts the model's generalizability. Future research should prioritize expanding data sources to integrate more diverse information from various child populations. Moreover, continuous exploration of other advanced algorithms and model combinations is essential to optimize the model's performance, enhance the precision and reliability of the assessment, and promote the scientific and accurate development of the field of children's mental health assessment. This research provides a solid foundation and valuable reference for subsequent studies and holds significant potential for practical application in promoting children's mental health.

References

[1] Wang Jiachuang, Dong Longjun. Risk assessment of rockburst using SMOTE oversampling and integration algorithms under GBDT framework [J].Journal of Central South University, 2024, 31(08): 2891-2915.

[2] Zhu Y, Chu Xilin, Le Yao. Differential dimensions of the distribution of dummy words in humanities and social science journal articles and their linguistic functions [J]. International Journal of Chinese Language Teaching, 2025, 6(1).

[3] Wang Peng, Deng Zhi, Fan Yuyuang. Anti-SEU Design for Finite State Machine with One-hot Encoding [J].Telecommunications Technology, 2022, 62(08): 1178-1183.

[4] FU Zhongliang, CHEN Xiaoqing, REN Wei, et al. Stochastic K nearest neighbor algorithm with learning process [J]. Journal of Jilin University (Engineering Edition), 2024, 54(01): 209-220.

[5] Zhu Yuanli, Feng Xiangyang, Yan Qingwu, et al. Study on spatial differentiation of soil organic carbon and main controlling factors in Northeast black soil area farmland based on gradient enhancement decision tree [J]. China Environmental Science, 2024(05): 1-14.