Predicting Mental Health Outcome of Middle-School Students

Zachk Huang

Dept. of Sociology

GitHub Repository: https://github.com/zachkhuang/zachkhuang.git

1 Introduction

This report investigates the prediction of students' mental health outcomes using a comprehensive set of features, including individual, family, and school-level factors. The primary objective is to enhance the identification of at-risk students and support timely, targeted interventions.

The analysis is based on data from the China Education Panel Survey (2013–2014), a rich dataset designed using a four-stage probability sampling method to ensure representativeness and randomness. This dataset provides detailed insights into the social and educational contexts of students. The predictive task is framed as a regression problem, with the target variable derived from responses to five mental health-related questions in the survey: "How often have you felt sad, blue, depressed, unhappy, or not enjoying life in the past seven days?" Each question is scored on a scale from 1 to 5. The target variable is calculated as the average score of these five measures, which is then reverse-coded so that higher scores indicate better mental health.

While no prior work has specifically predicted mental health outcomes using this dataset, there is extensive research on predicting students' mental health. Most of these studies, however, focus on classification tasks (Shafiee and Mutalib, 2020). For example, Sahlan et al. (2021) predicts mental health problems among university students, achieving an accuracy score of 0.64. Other studies, such as Sahu et al. (2023), report higher performance metrics (e.g., a precision of 0.95 and recall of 0.86) by incorporating a series of subjective measures of mental well-being as features in their data.

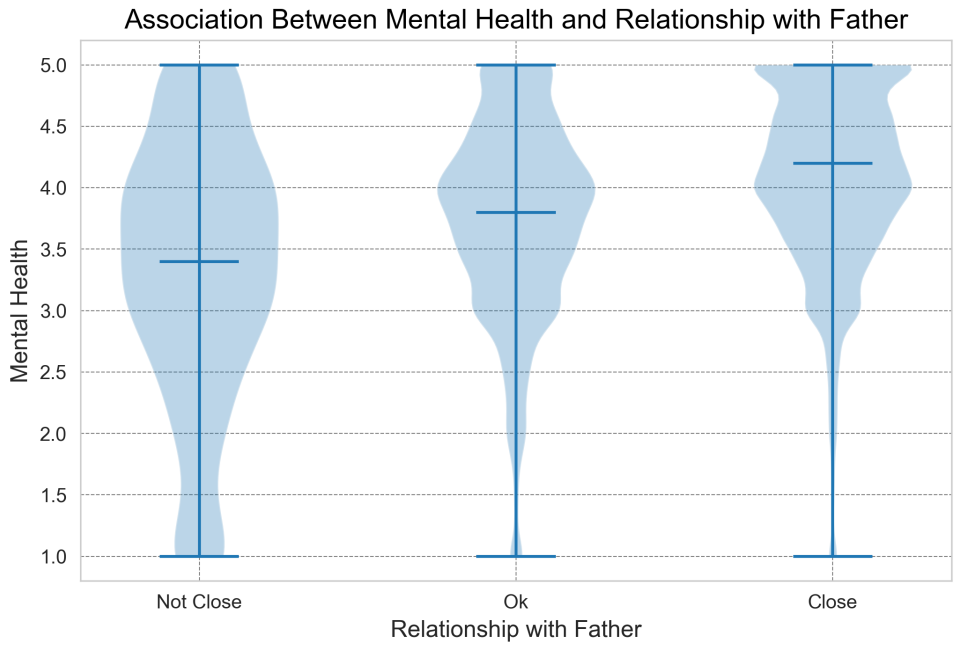

Figure 1: Violin plot showing the association between relationship with father and mental health

2 Explanatory Data Analysis

In this section, I will present plots that illustrate the relationship between the target variable and three key features: students' relationship with their fathers, family socioeconomic status, and whether parents quarrel.

Figure 1 depicts the association between the relationship with the father and mental health, shown using a violin plot. The relationship with the father is an ordinal feature with three levels: "not close," "okay," and "close." Students with a closer relationship with their fathers tend to exhibit better mental health: their median mental health score is higher, and the distribution is more concentrated in the upper half of the range.

Figure 2 presents category-specific histograms of mental health, grouped by whether parents quarrel with each other. The results suggest that students whose parents are on better terms generally experience better mental health on average. In contrast, students with quarreling parents tend to have poorer mental health.

Figure 3 illustrates how mental health varies across family socioeconomic status, measured by students' families' current financial situation, using box plots. The results reveal a clear socioeconomic gradient in mental health among middle school students, with those from more affluent families generally exhibiting better mental health.

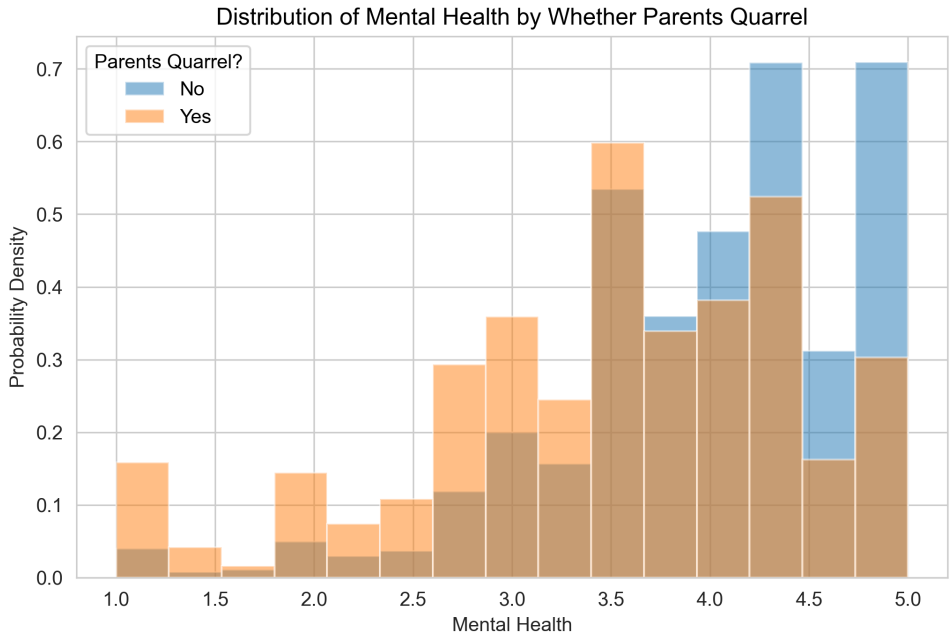

Figure 2: Category-specific histogram showing the association between whether parents quarrel and mental health

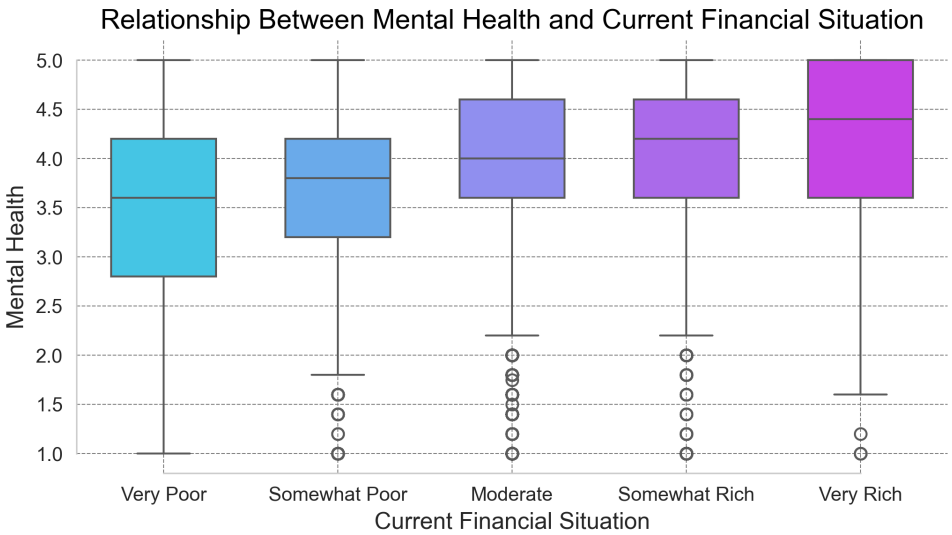

Figure 3: Boxplot showing the association between current financial situation and mental health

3 Methods

3.1 Data Splitting

Since there is no group or time-series structure and the target variable is continuous, I use a simple splitting strategy to divide the data into a training set, validation set, and test set. Specifically, I first split the dataset into X_test, X_other, y_test, and y_other. Then, I further split the X_other and y_other into a training set and validation set. The training set comprises 60% of the data, while the validation and test sets each account for 20

I chose not to use K-Fold cross-validation because I am employing a pattern submodel approach to handle missing values in the continuous features. Since there are over 20 different missing patterns, K-Fold cross-validation would significantly increase the number of models to train, making it computationally expensive.

3.2 Preprocessing

For categorical features in my dataset, I use the OneHotEncoder to transform them into dummy variables and the SimpleImputer to handle missing values. For ordinal features, I apply the OrdinalEncoder to convert the variables into ordered numerical values, and again use SimpleImputer for imputation. For continuous features, I use the StandardScaler to standardize the data and do not impute the missing values. Before preprocessing, the dataset contains 120 features; after preprocessing, this number increases to 268 features.

3.3 Machine Learning Algorithms and Hyperparameter Tuning

Table 1: Summary of Machine Learning Algorithms

Algorithm	**Hyperparameters**	**Missing Values**
Lasso Regression	alpha:np.logspace(-4, 1, 6)	Reduced-features models
Elastic Net	alpha:np.logspace(-4, 1, 6) l1_ratio: np.linspace(0.01,1,6)	Reduced-features models
Random Forest	max_depth:[1, 3, 10, 30, 100] max_features: [0.25, 0.5,0.75,1.0]	Reduced-features models
XGBoost	reg_alpha:[0e0, 1e-2, 1e-1, 1e0, 1e1, 1e2] reg_lambda:[0e0, 1e-2, 1e-1, 1e0, 1e1, 1e2] max_depth :[1,3,10,30,100]	Handled directly

Table 1 summarizes the four machine learning models I trained, the hyperparameters I tuned, the corresponding hyperparameter grids, and how each model handles missing

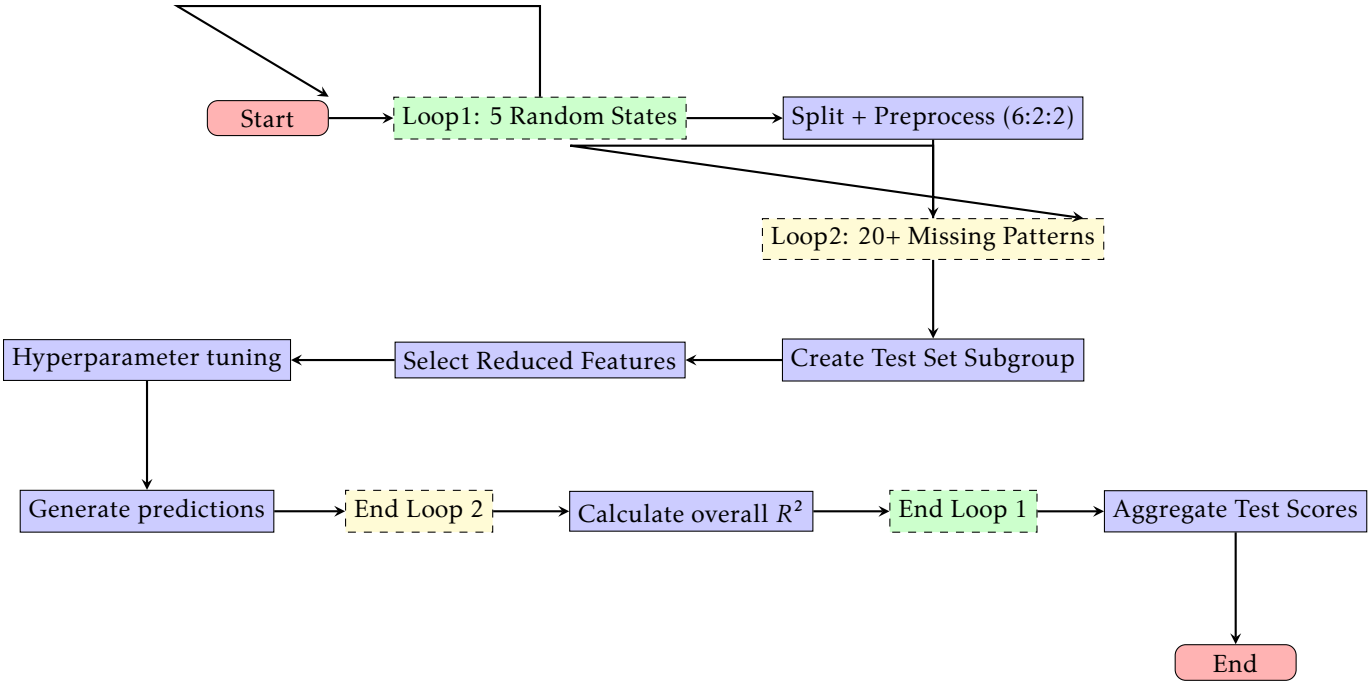

Figure 4: Machine Learning Pipeline with Reduced-Features Model

values in continuous features. The hyperparameters with no upper bound are spaced evenly on a logarithmic scale, while those with clear lower and upper bounds are spaced linearly. I expanded the range of the hyperparameters until the optimal parameters were no longer at the edge of the range.

Since Lasso Regression, Elastic Net, and Random Forest models cannot handle missing values by themselves, I used a reduced-features model approach to address this issue. For the XGBoost model, I did not apply this approach.

In general, an alpha of 0.001 yields the best results for both Lasso Regression and Elastic Net. The best value for the l1_ratio in Elastic Net varies across different models I trained. For the Random Forest model, the best hyperparameters differ significantly across the various pattern submodels. For XGBoost, a max_depth of 3 generally produces the best results, although the optimal values for the other two hyperparameters vary across different random states.

3.4 Machine Learning Pipeline

Figure 4 illustrates the machine learning pipeline with the reduced-features model for the first three algorithms. I use R^2 as the evaluation metric because it is straightforward and easy to interpret.

The process begins by creating 5 random states and looping through each of them. For

each random state, I first split and preprocess the dataset. Then, I identify more than 20 missing patterns in the test set and loop through each unique missing pattern. Within each pattern, I create a test set subgroup, select reduced features based on the missing values, and identify the data points in the training and validation sets that do not contain missing values in those features.

Next, I train the model and tune the hyperparameters based on performance in the validation set. Once the best hyperparameters are identified, I generate predictions for the test set subgroup corresponding to that missing pattern. After looping through all the missing patterns, I obtain predictions for all data points in the test set, allowing me to calculate the overall R^2 for the test set. Once I have the overall R^2 scores for all 5 random states, I aggregate them by calculating the mean and standard deviation. The standard deviation serves as a measure of uncertainty in the test scores.

For the XGBoost model, the pipeline omits the pattern submodel part (i.e., the second loop).

4 Results

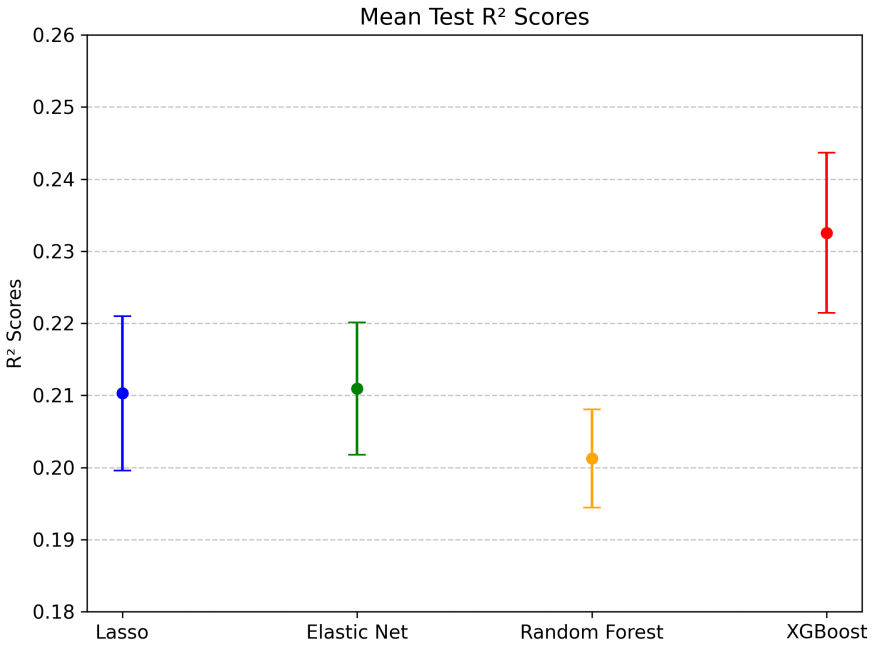

Figure 5: Test R^2 Scores Across 4 Models

Figure 5 shows the performance of the four machine learning algorithms. The baseline

test score (R^2) is 0, which corresponds to using the average mental health score as the prediction for all data points. This serves as a reference for comparison.

XGBoost achieves the best performance, with a mean test R^2 of 0.2325 and a standard deviation of 0.0111. This result is more than 20 standard deviations above the baseline score, indicating a significant improvement over the baseline. Furthermore, XGBoost outperforms the other three algorithms by a considerable margin.

To interpret and understand the model better, I calculated three types of global feature importance, as well as local feature importance for two data points in my sample, using the best XGBoost model from the last random state.

Figure 6 shows the ten features with the highest permutation feature importance. Students' current health is the most predictive factor. Other strong predictors of mental health include the level of stress students feel about their parents' expectations, parents' confidence in their children, and average sleeping and studying time. Additionally, features such as whom students turn to for help and when they want to chat, their relationship with their father, whether they had an illness in preschool, and their math score also rank in the top 10 most important features.

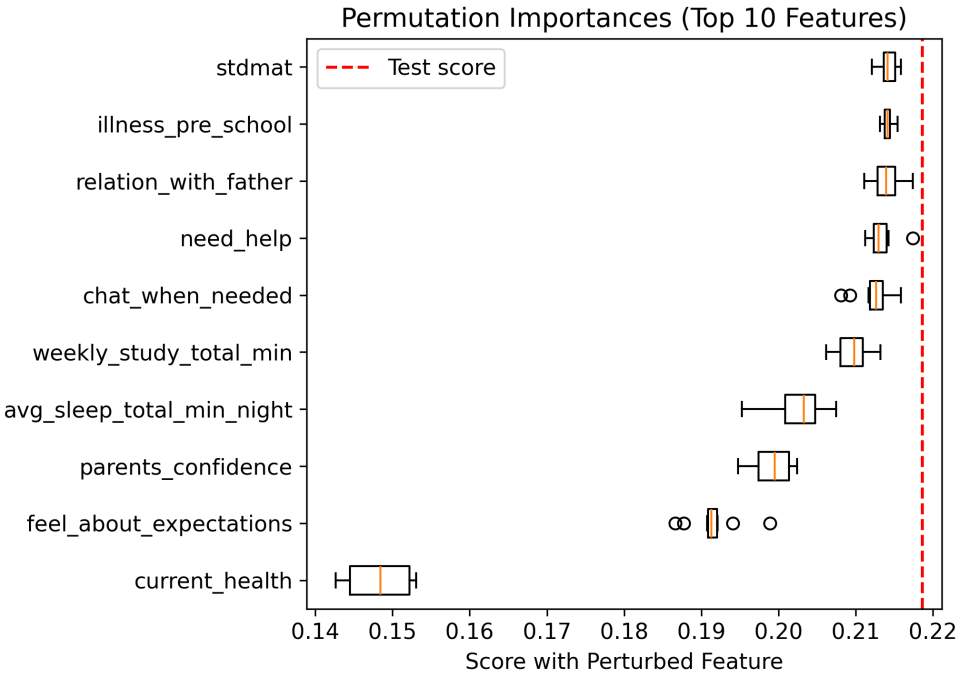

Figure 6: Global Permutation Feature Importance (Top 10)

Figure 7 shows the ten features with the highest total_gain in the XGBoost model, which measures the cumulative improvement in model performance from each feature, reflecting its overall contribution to reducing the error across all splits. Nine of the top

10 features overlap with those identified by permutation importance, reinforcing the robustness of these predictors. Students' current health, parental confidence, sleeping time, feelings about expectations, and study time continue to be strong predictors of mental health. Notably, the features "whom students turn to when they want to chat" and "relationship with their father" become more prominent when assessed using total_gain.

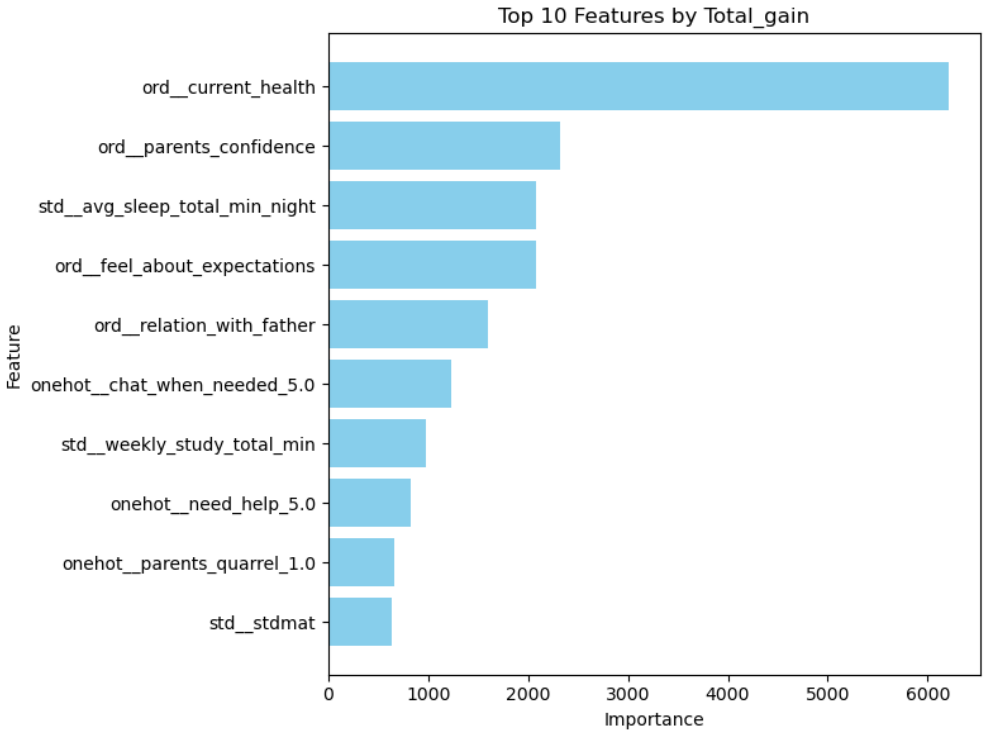

Figure 7: Global Feature Importance: Total Gain (Top 10)

Figure 8 shows the ten features with the highest mean absolute SHAP value. Once again, nine out of the top 10 features overlap with those identified by permutation feature importance. Current health remains the most powerful predictor, with the other top predictors also showing strong consistency. Overall, the global feature importance remains highly consistent across different metrics.

Figure 9 and Figure 10 show the SHAP force plots for two data points in my test set. For the first student, whose mental health prediction is lower than the baseline, her parents are very confident in her. She has a good math score, a strong relationship with her father, and spends less time studying—all of which contribute positively to her mental health prediction. However, she has moderate health, appears to experience sleep problems, and feels stressed about her parents' expectations, all of which contribute negatively to her prediction.

For the second student, her mental health prediction is higher because she has good

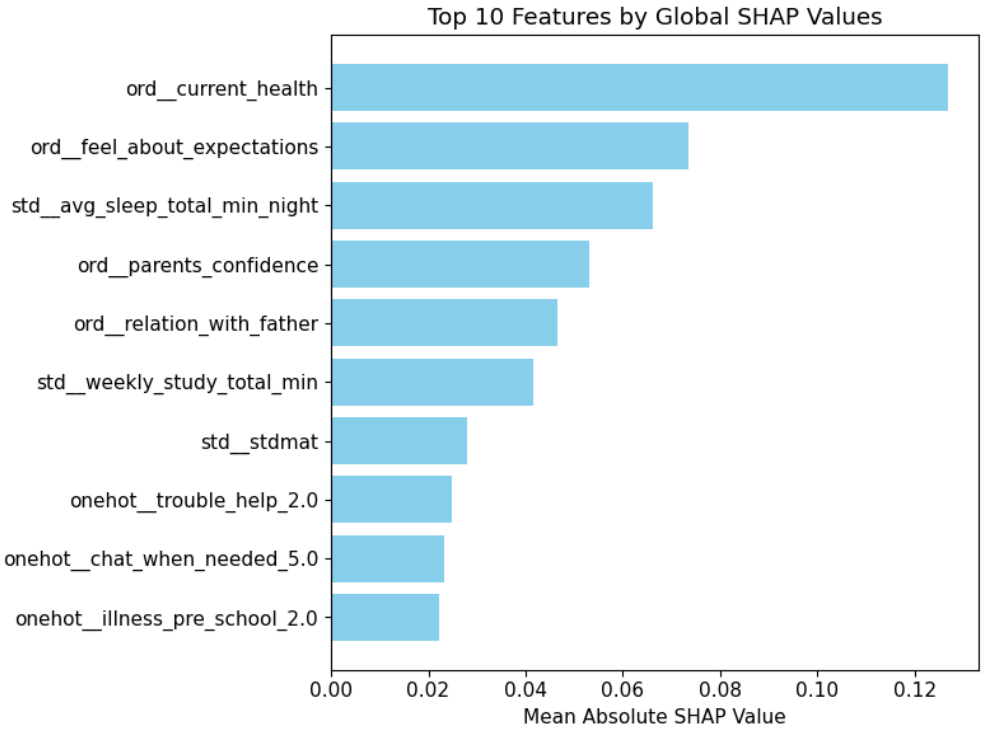

Figure 8: Global Feature Importance: Mean Absolute SHAP Value (Top 10)

health, does not feel stressed about her parents' expectations, and has a good math score. However, she has lower sleeping time, spends a lot of time studying, and has a moderate relationship with her parents, all of which negatively impact her prediction.

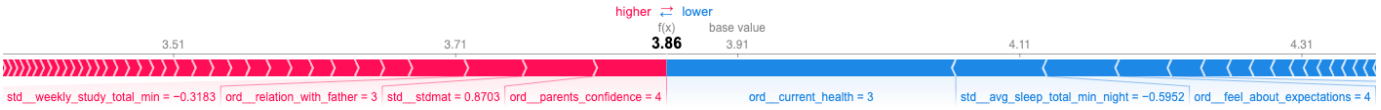

Figure 9: SHAP Force Plot 1: index 19

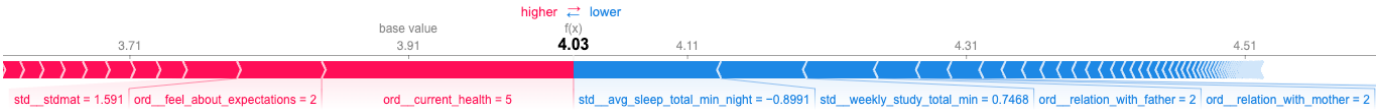

Figure 10: SHAP Force Plot 2: index 36

Generally speaking, the strong predictors identified from both global and local feature importance align with real-life expectations and our common understanding. These predictors also demonstrate strong consistency and robustness across different metrics, highlighting the model's good interpretability.

5 Outlook

The model's test scores are not high, but this is expected given that the data comes from social science. Much of this can be attributed to the fact that mental health can be influenced by almost everything in one's life, and there are numerous unobserved factors that cannot be captured by survey data. Additionally, the mental health measure is highly unstable, as it can be significantly impacted by events that occurred shortly before the survey was filled out. For example, the student might have just been scolded by a teacher or had a small argument with a friend.

Given these challenges, improving the model's predictive power will require more sophisticated approaches. Exploring deep learning models and tuning additional hyperparameters could be effective strategies. Additionally, using K-fold cross-validation would help make better use of the data and reduce bias. We could also consider predicting each of the five individual measures rather than using the average score, or transforming the target variable into a binary indicator (e.g., whether a student needs intervention) for better practicality in deployment. However, to truly enhance the model's performance, collecting more comprehensive data, such as information on students' interactions with their peers and friends, would be essential.

References

(2013). China education panel survey: http://ceps.ruc.edu.cn/english/overview/overview.htm.

Sahlan, F., Hamidi, F., Misrat, M. Z., Adli, M. H., Wani, S., and Gulzar, Y. (2021). Prediction of mental health among university students. *International Journal on Perceptive and Cognitive Computing*, 7(1):85–91.

Sahu, B., Kedia, J., Ranjan, V., Mahaptra, B. P., and Dehuri, S. (2023). Mental health prediction in students using data mining techniques. *The Open Bioinformatics Journal*, 16(1).

Shafiee, N. S. M. and Mutalib, S. (2020). Prediction of mental health problems among higher education student using machine learning. *International Journal of Education and Management Engineering*, 10(6):1.