Student Stress/Depression Level Monitoring

Summary:

This project leverages **machine learning** to predict and analyze student stress levels using psychological, academic, and lifestyle factors. Through **data preprocessing, feature selection, and advanced models** such as Logistic Regression, Random Forest, and XGBoost optimized with Bayesian techniques, the system identifies key stress indicators and classifies students into **low, medium, or high stress categories**. In addition, the project extends to predicting a student's **PHQ-9 score** based on survey responses, providing deeper insights into depression risk. The findings highlight the most influential factors affecting student well-being and can guide interventions to improve mental health support.

Project Goal and Machine Learning System

- **Project Goal:** The primary goal was to utilize machine learning to predict and analyze student stress levels by combining psychological, academic, and lifestyle factors. The system was designed to classify students into low, medium, or high stress categories.
- **Extension:** The project also extended this analysis to predict a student's **PHQ-9 score**, providing deeper insights into depression risk based on survey responses. The overall findings were intended to highlight the most influential factors affecting student well-being, guiding potential interventions.

Data Source and Initial Exploratory Data Analysis (EDA)

Before building the models, it was important to first explore the dataset. This step made it possible to understand how each factor varied among students while also checking for missing values, outliers, or imbalances. **Key Data Features:** The dataset included features related to mental health and lifestyle, such as:

- `anxiety_level`
- `self_esteem`
- `sleep_quality`
- `depression`
- `academic_performance`
- `bullying`
- `blood_pressure`
- `safety`

Data Quality Confirmation:

- By examining these patterns, I ensured the dataset was clean and reliable, and I could identify which features were most likely to play a key role in predicting student stress and depression levels.
- A statistical summary highlighted averages, ranges, and variability across features. For example, 1100 instances were analyzed, with the mean `anxiety_level` being approximately 11.06, and the mean `depression` score being 12.55.

Handling Outliers and Features:

- Boxplots identified outliers in `Noise Level` and `Living Conditions`.
- However, these outliers were **retained** because the features were measured on a fixed scale (1–5), meaning the extreme values were considered valid, meaningful responses that could contribute important information to stress and depression modeling.

Target Variable (Stress) Understanding:

- The target variable, `stress_level`, included three categories: **0 = No Stress, 1 = Eustress, and 2 = Distress**.
- To better understand how the features relate to one another, I then computed the **Pearson correlation matrix**. This allowed me to see which factors were strongly or weakly correlated, both with each other and with the target variable. Identifying these relationships was important because it highlighted potential predictors of student stress, and depression levels.

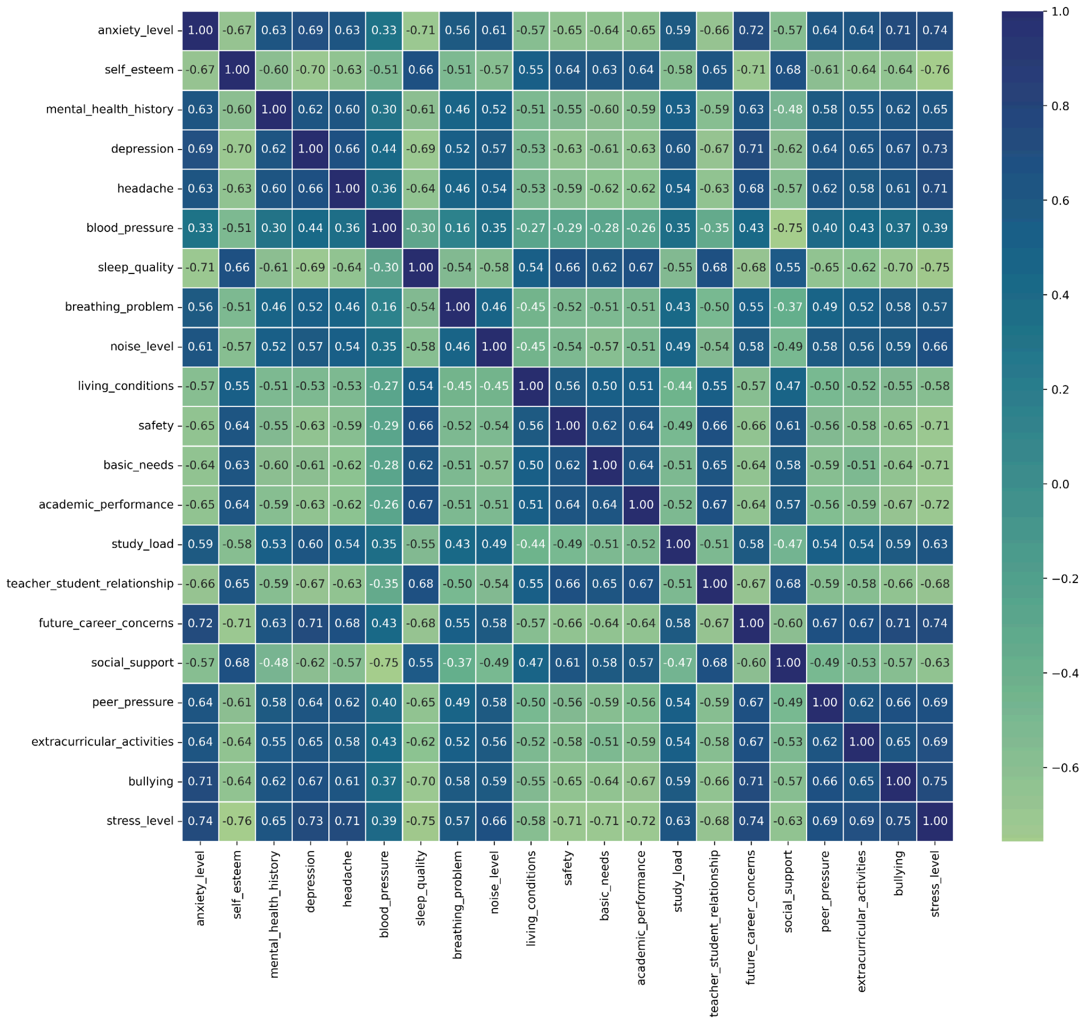

Fig 1. Pearson Correlation Matrix

Machine Learning Model Selection

Three distinct machine learning models were chosen to predict student stress levels, ensuring a robust comparison:

Model	Rationale for Selection

Logistic Regression	Chosen as a good baseline model for comparison against more complex methods, it works well for categorical target variables (stress level).
Random Forest	Handles non-linear relationships; robust to noise and outliers; provides **feature importance** scores.
XGBoost	Excellent for handling structured/tabular data; efficiently handles missing values and outliers; includes built-in regularization to prevent overfitting; also provides feature importance scores for interpretability.

Table 1. Machine Learning Models and Rationale for Selection

Model Performance and Comparison (Stress Level)

All three models demonstrated strong performance in classifying student stress levels:

- **Logistic Regression** achieved the highest overall accuracy (90.9%).
- **Random Forest** followed closely (90.0%).
- **XGBoost** performed robustly (89.5%).

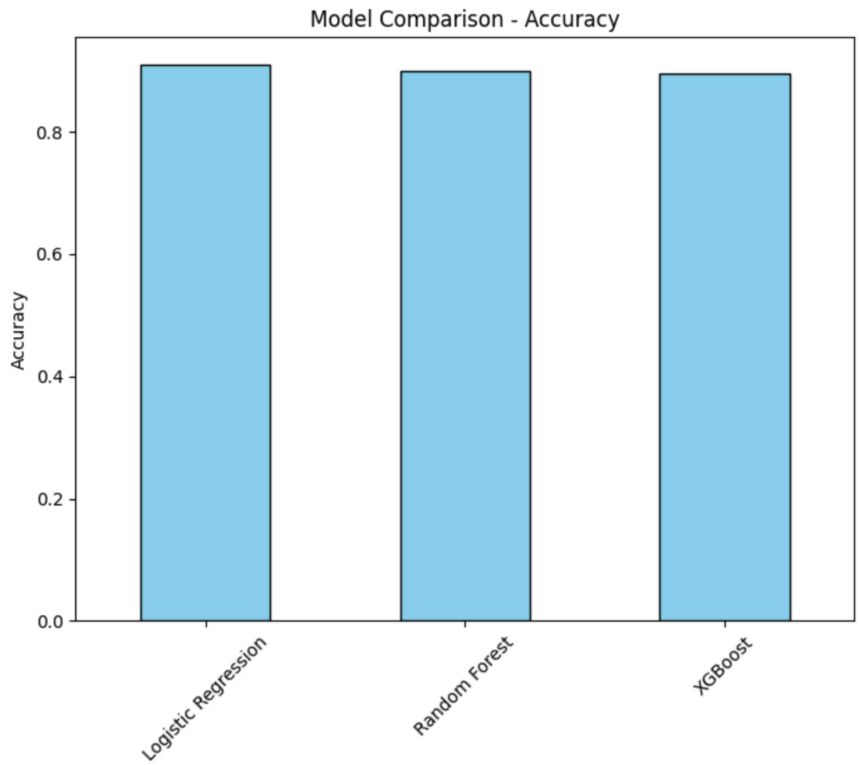

Fig 2. Model Comparisons (Accuracy Metric)

Logistic Regression slightly outperformed the others in both precision and recall, making it the most effective predictor for this specific dataset for the stress level.

Deep Dive: Logistic Regression Performance

To illustrate the model's effectiveness, the classification report for Logistic Regression revealed specific performance across the three categories:

- **Class 0 (No Stress):** Demonstrated **high precision (0.97)**, meaning when the model predicted 'no stress,' it was very accurate. It had slightly lower recall (0.84).
- **Class 1 (Eustress):** Showed **very high recall (0.97)**, meaning the model captured almost all actual eustress cases, although it sometimes confused them with other stress levels (lower precision of 0.84).
- **Class 2 (Distress):** Achieved balanced performance (0.94 precision, 0.92 recall).

Macro and Weighted Averages (both 0.91–0.92) confirmed that the model was consistently effective across all classes, without being biased toward the majority category.

Key Stress Predictors

Analyzing feature importance across all models provided crucial insights into factors influencing student stress. Variables that **consistently emerged as key predictors** were:

1. **Sleep Quality**
2. **Safety**
3. **Basic Needs**
4. **Bullying**

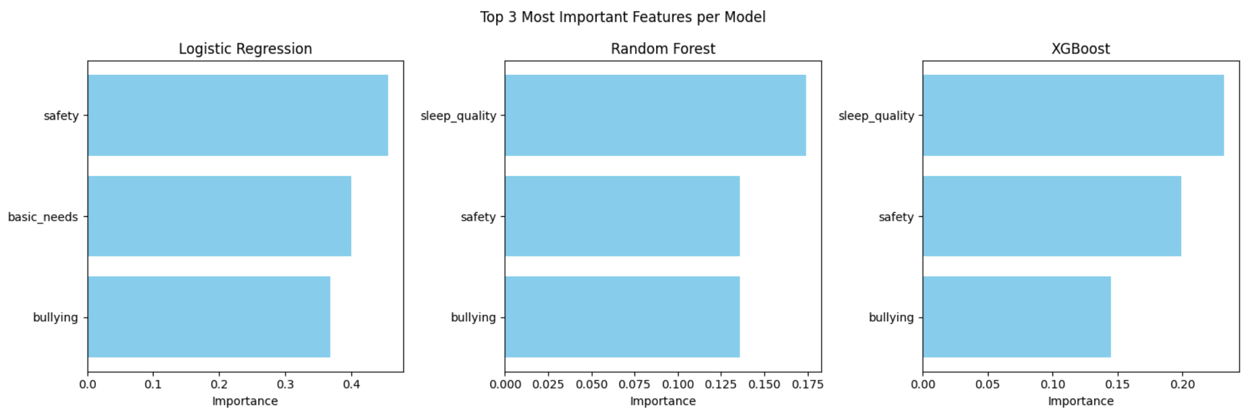

Fig 3. Top 3 Most Important Features per Model

This highlighted that both personal well-being (sleep, basic needs) and environmental factors (safety, bullying) significantly influence student stress levels.

The Challenge: Imbalanced Depression Data

Motivated by the presence of the PHQ-9-scored ‘Depression’ column, a second phase of this project is aimed at predicting the severity of mental health.

The dataset for depression severity was found to be **imbalanced**. Specifically, the **‘Moderately Severe’ class had the fewest samples** (around 150), while the ‘Moderate’ class had significantly more (over 300).

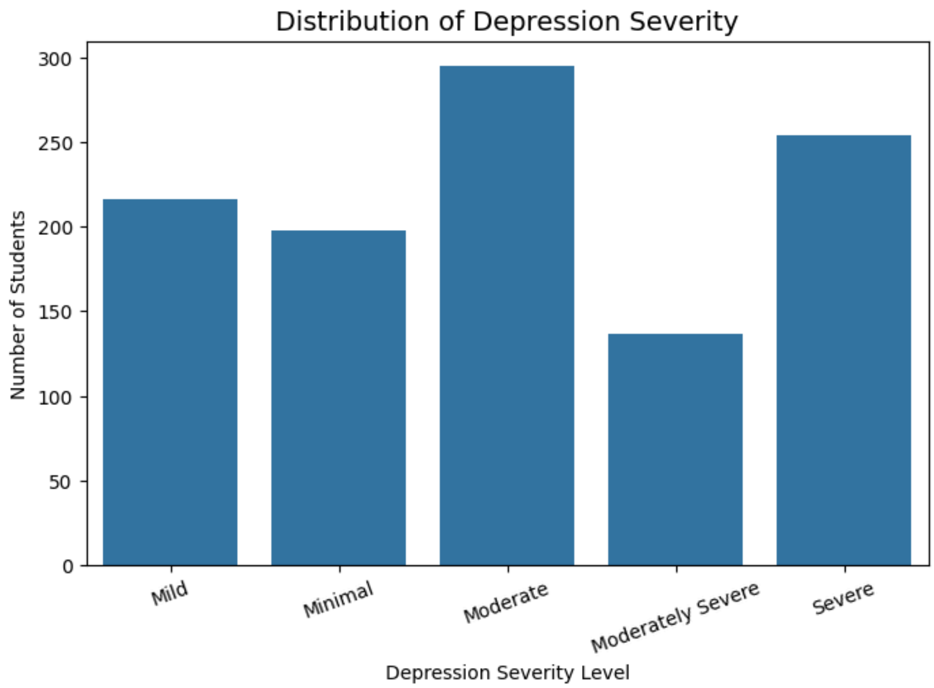

Fig 4. Distribution of Classes of Depression Severity

This imbalance directly impacted the initial model attempts:

- The initial logistic regression achieved limited effectiveness, with an accuracy of only **59%**.

Techniques implemented for Handling Imbalance and Tuning

To address the poor initial performance and the inherent data imbalance, several techniques were systematically applied and tested:

- **Bayesian Optimization:** Used to automatically search for and select the best model hyperparameters, making the tuning process more efficient than random or exhaustive searches.
- **SMOTE (Synthetic Minority Oversampling Technique):** Used to generate synthetic samples for minority classes by interpolating between real samples, thereby creating a more balanced training dataset.
- **Class Weights:** Applied to penalize the model more heavily for misclassifying underrepresented (minority) classes during training, forcing the model to pay more attention to these categories.
- **Stratified K-Fold Cross-Validation:** Implemented to ensure that the class distribution remained consistent across all training folds, aiding the model in learning across all severity categories.

Performance Results for Depression Prediction

Despite implementing sophisticated techniques, the prediction accuracy for depression severity remained challenging, with a plateau of 56% to 61% across all models and configurations.

Model Variant	Overall Accuracy	Key Finding/Trade-off
LogReg (SMOTE + Bayes)	56–61%	Performance plateaued, suggesting fundamental limits with the current features.
Random Forest (Weighted)	52%	Successfully predicted samples from the previously missed ‘Moderately Severe’ class, but at the cost of overall accuracy due to the shift in focus towards balancing the classes.
XGBoost (Weighted)	56%	Showed better consistency in predicting Mild and Moderately Severe classes compared to the baseline, improving the F1-scores for these underrepresented categories.

Table 2. Performance and Key Findings of Model Variants

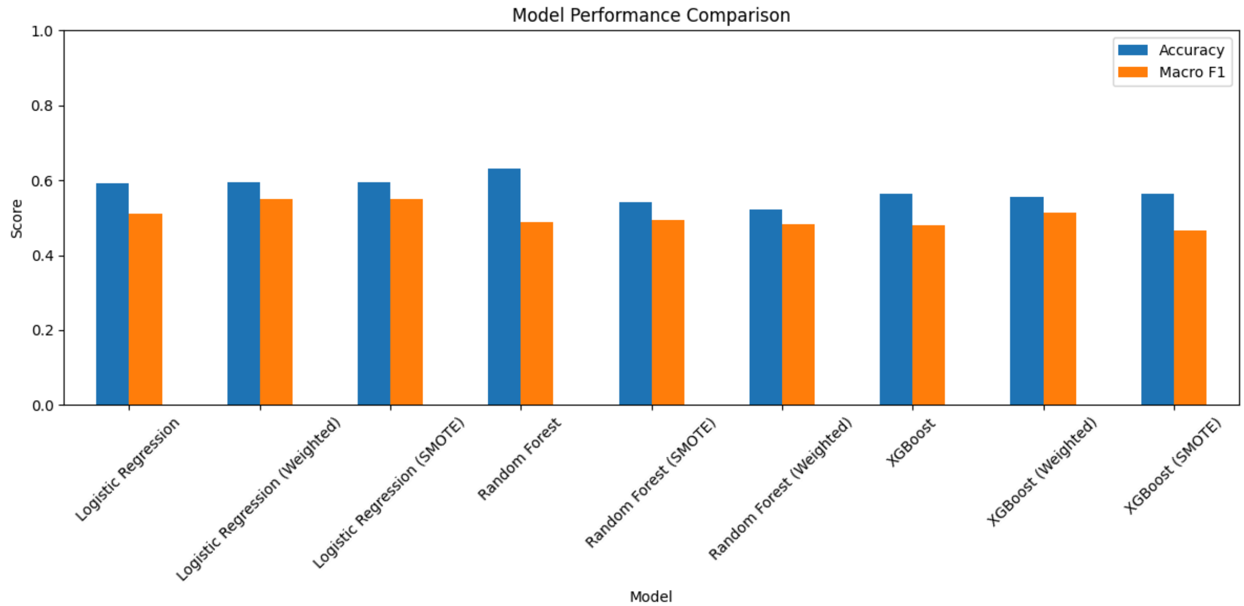

Fig 5. Graph for Model Performance for Depression Severity (Accuracy, Macro F1)

Evaluation Conclusion: All models delivered similar overall performance. While Random Forest achieved the highest accuracy in some trials, it often had a lower macro F1-score (harmonic mean of precision and recall), indicating a weaker balance across classes despite using techniques to fix imbalance. The Weighted and SMOTE variants of Logistic Regression demonstrated **more consistent accuracy and macro F1 scores**, making them more dependable choices for real-world applications where avoiding bias against minority classes is crucial.

Key Depression Predictors

Analyzing feature importance across all models provided crucial insights into factors influencing student depression. Variables that **consistently emerged as key predictors** were: The variable that consistently emerged as a key predictor is `blood_pressure`.

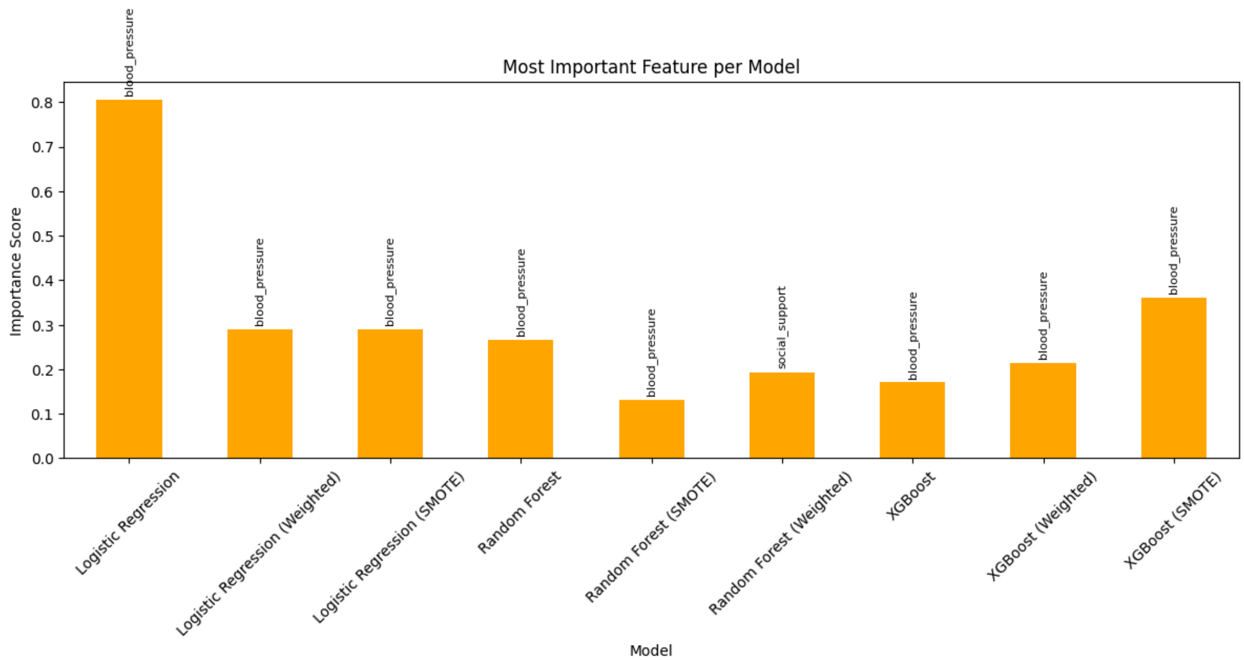

Fig 6. Most Important Predictor for Depression Severity

Final Conclusions and Limitations

Stress Prediction (Success): The project successfully built robust models (Logistic Regression, Random Forest, XGBoost) capable of predicting student stress levels with high accuracy (around 90%). Key predictors included `sleep quality`, `safety`, `basic needs`, and `bullying`.

Depression Prediction (Challenge): Predicting depression severity proved significantly more challenging due to severe class imbalance. Even after applying advanced techniques like Bayesian Optimization, SMOTE, and class weights, the models struggled to generalize across all severity levels, achieving only moderate accuracy (56–61%).

Most Influential Feature for Depression: The analysis consistently revealed that **blood pressure emerged as the top feature** across nearly all models when attempting to predict Depression severity.

Future Development Recommendations

The major limitation in depression prediction was the predictive power of the existing feature set. For future development, it is highly recommended to **incorporate additional, more relevant features** to improve the model's ability to accurately predict depression levels.

Kaggle Dataset:
https://www.kaggle.com/datasets/mdsultanulislamovi/student-stress-monitoring-datasets/data