# Student Stress Level Prediction Model

- *Tanvi Joshi, 15 Oct 2025*
- DataSet
- To predict student stress levels (Low, Medium, High) using behavioral, physical, and environmental data.
- To create an accurate and interpretable predictive tool that can serve as an early warning system for educational institutions.

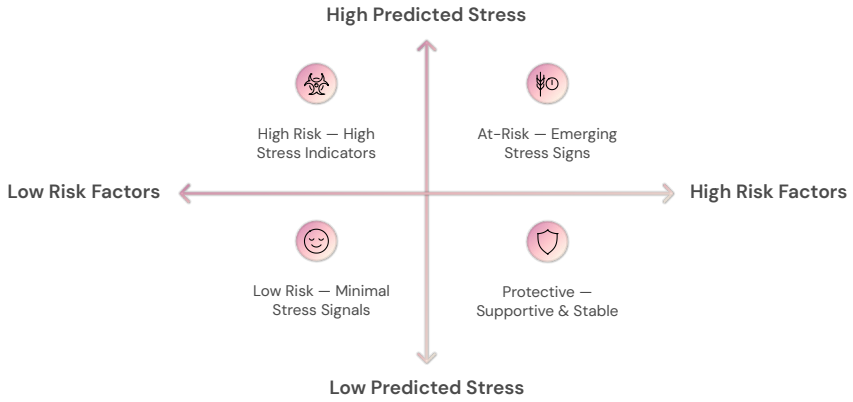
# The Challenge: Predicting Student Stress Levels

## What is the Problem?

Educational institutions face a critical challenge—identifying students experiencing stress before it leads to serious mental health issues. This project aims to ask:

*Can we predict a student's stress level (Low, Medium, High) by understanding the behavioral, physical, and environmental factors that affect them?*

- **Dataset:** 1,100 students, 20 variables, 1 target variable
- **Target classes:** Low (373), Medium (358), High (369) – balanced distribution. **No class imbalance handling** needed.
- **Goal:** Build an accurate, interpretable model
- **Purpose:** Serve as an early warning system for proactive mental health support




**85%**

**Target Accuracy**

Minimum threshold for clinical reliability

**90.9%**

**Achieved Result**

From Tuned Random Forest Model

Final model performance exceeded expectations

**20**

**Predictor Variables**

Behavioral, physical, and environmental factors such as blood pressure, anxiety level, self esteem, academic performance etc.

# Methodology: A Supervised Learning Approach

### Data Split & Stratification

- **Split Ratio:** 80% training, 20% testing
- **Stratification:** Used to maintain class balance in both sets
- **Random State:** Fixed at 42 for reproducibility
- **Result:** 880 training samples, 220 testing samples
- **StandardScaler** was applied for Logistic Regression, and **5-fold cross-validation** ensured model stability.

### Evaluation Metrics

- **Accuracy:** Overall correct predictions
- **Precision:** When we predict a stress level, how often are we correct?
- **Recall:** Of all students at each stress level, how many did we identify?
- **F1-Score:** Harmonic mean of precision and recall
- **Confusion Matrix:** Detailed breakdown of predictions vs actual values

```
MODEL COMPARISON SUMMARY
Logistic Regression Accuracy: 0.882 (88.2%)
Random Forest Accuracy:       0.891 (89.1%)
XGBoost Accuracy:             0.873 (87.3%)
```

**Workflow:** *Data → Preprocessing → Model Comparison → Hyperparameter Tuning → Deployment.*

1

#### Data Selection Through EDA

Selected StressLevelDataset.csv containing 1,100 student records with 20 clean, numeric features. The dataset exhibits excellent balanced class distribution, ensuring fair and robust model training across all stress categories.

2

#### Model Comparison

Compared Logistic Regression, XGBoost, and Random Forest to identify the best model. RF was selected for its feature interpretability, ability to capture non-linear relationships, and low variance in 5-Fold CV (stability). Its 0.9% accuracy gain over Logistic Regression (89.1% vs. 88.2%) was not statistically significant ($p \geq 0.05$, McNemar's Test).

3

#### Hyperparameter Tuning

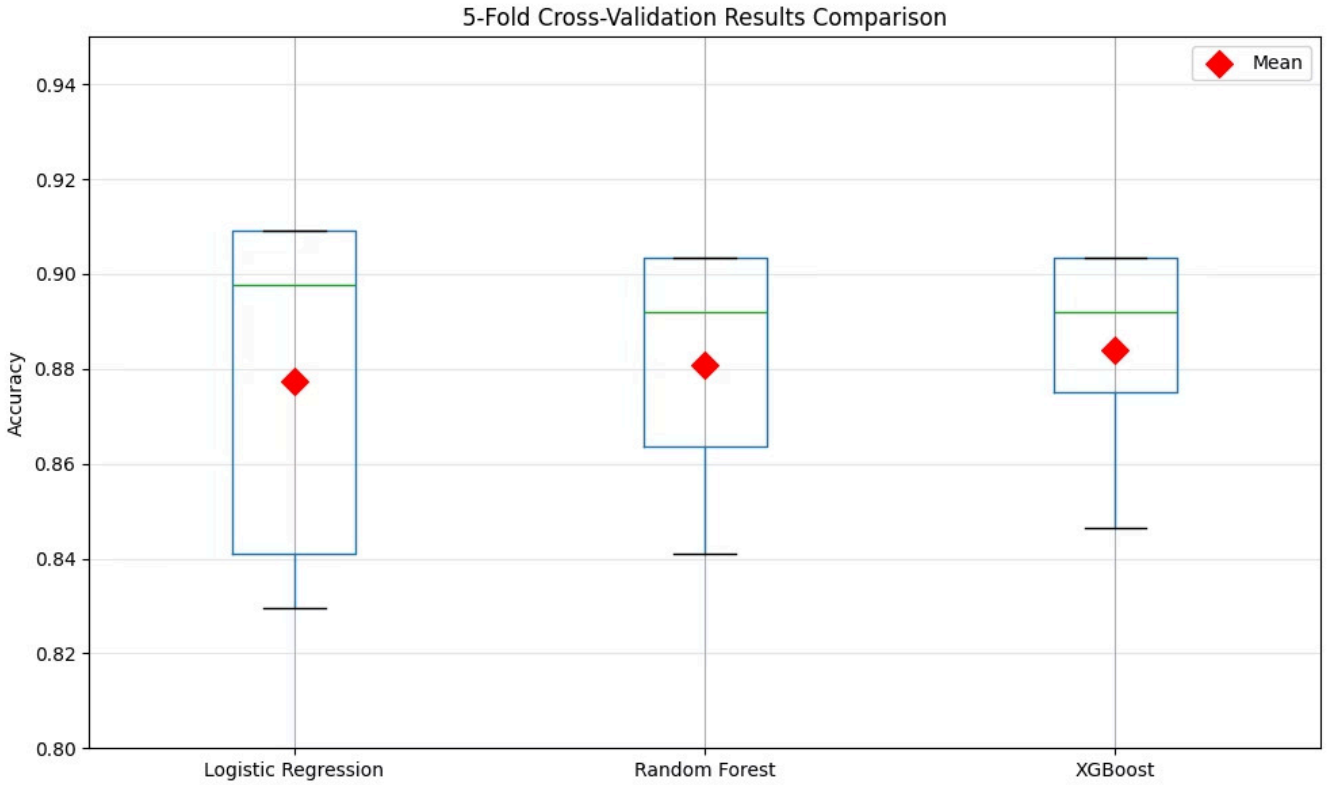
Implemented GridSearchCV with 5-fold cross-validation to systematically tune model parameters, preventing overfitting whilst maximizing generalization to unseen student data.

yielded a final accuracy of **90.9%**—a **1.82% gain** over the default model.

4

#### Deployment/ Actionable Intelligence

The resulting model provides detailed feature importance rankings, translating predictions into practical, targeted intervention strategies for school mental health coordinators.




## CROSS-VALIDATION CONCLUSION

CROSS-VALIDATION CONCLUSION

Random Forest shows consistent performance (std: 0.025)
Low standard deviation indicates the model is stable across different data splits.

**Implications for Model Selection:**

- Even if not statistically significant, Random Forest provides feature importance
- Random Forest handles non-linear relationships better
- Marginal accuracy gain + interpretability = justified choice

# Key Factors Influencing Student Stress

The exploratory data analysis examined correlations between all features and the target variable, revealing both protective factors that reduce stress and risk factors that increase vulnerability. The final Random Forest model provided definitive feature importance rankings.

**Physiological**

Blood Pressure

**Type:** Risk Factor

**Initial Correlation:** Complex (non-linear, strongest predictor)

**Final Importance:** **0.157479**. (Rank 1)

The single most powerful predictor, indicating that stress manifests physically before psychological awareness develops (double that of the next factor.

**Health Behavior**

Sleep Quality

**Type:** Protective Factor

**Initial Correlation:** r ≈ -0.75

**Final Importance:** **0.079318**. (Rank 2)

Poor sleep quality emerges as a critical warning signal for elevated stress levels.

**Academic**

Academic Performance

**Type:** Protective Factor

**Initial Correlation:** r ≈ -0.72

**Final Importance:** **0.074319** (Rank 3)

Strong academic performance, consistent achievement and academic confidence help buffer psychological strain.

**Psychological Factor**

Self-Esteem

**Type:** Protective Factor

**Initial Correlation:** r ≈ -0.76 (Strongest Negative Correlation)

**Final Importance:** **0.049751** (Rank 9)

High self-esteem provides significant psychological buffering. Interestingly, while showing the strongest initial negative correlation (r≈-0.76) , it ranks 9th in the final model importance (0.049), demonstrating how the random forest model analyzes feature interactions rather than just direct correlation.

**Social Factor**

Bullying

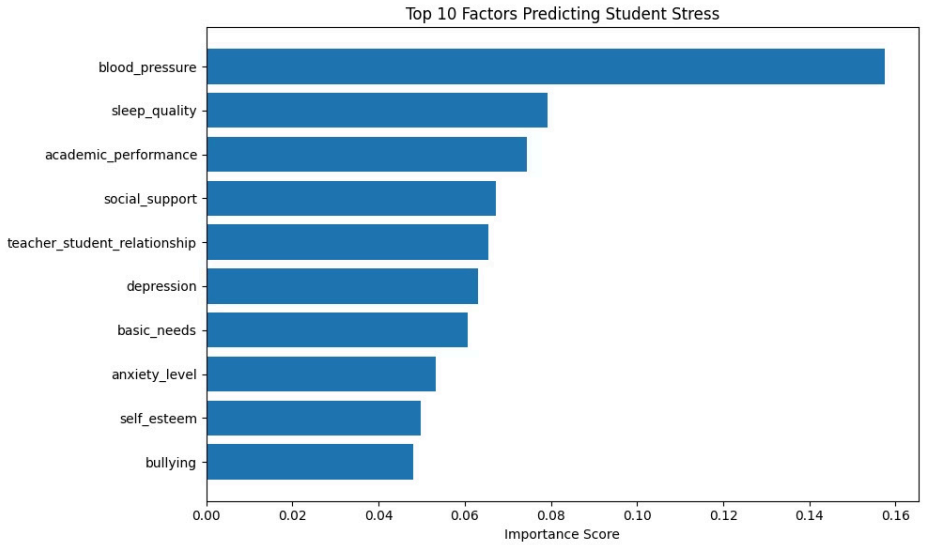
**Type:** Risk Factor

**Initial Correlation:** r ≈ +0.75 (Strongest Positive Correlation)

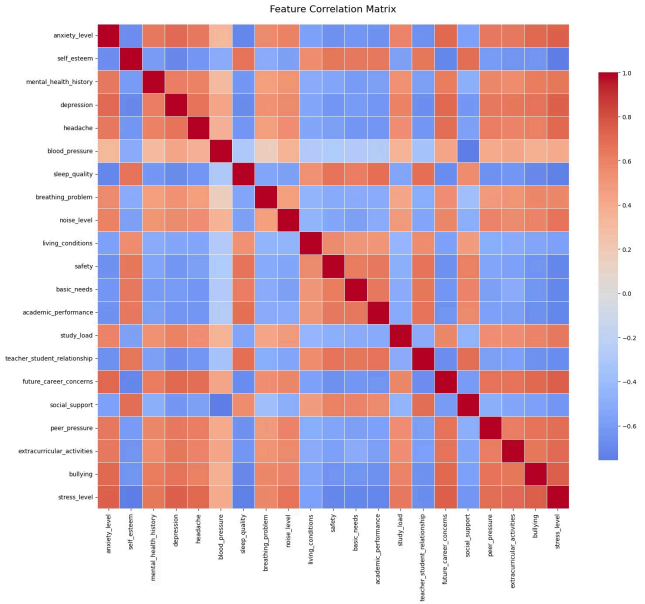
**Final Importance:** **0.048120** (Rank 10)

Bullying experiences strongly correlate with elevated stress

# Graphs




- **Key Insight :**
  - **The most powerful predictor is Blood Pressure**.
  - This is a **physiological marker**, not a traditional psychological indicator, suggesting that stress manifests physically first.
  - *Self-Esteem* and *Bullying* round out the bottom of the top 10.




**Key Insight :**

- **Validate Feature Relationships:** The heatmap shows **complex, non-linear relationships** between the 20 variables (e.g., between Bullying/Stress and Self-Esteem/Stress).
- **Supports RF Selection:** This visualization justifies using an ensemble method like Random Forest, which is **superior** to linear models for capturing these complex patterns.

# Model Performance and Configuration

## What Worked for Me?

The **Random Forest Classifier** emerged as the superior algorithm, successfully capturing the complex, non-linear relationships between the 20 predictor variables. Unlike simpler linear models, Random Forest's ensemble approach—combining multiple decision trees—proved essential for handling the multidimensional nature of student stress.

Hyperparameter tuning delivered a crucial 1.82% performance improvement over the default model configuration, demonstrating that systematic optimization is vital for achieving clinical-grade accuracy in educational mental health applications.

---

## Final Model Configuration

### Model Selection

Tuned Random Forest Classifier

### Accuracy Achieved

**90.9%** classification accuracy. This is a **1.82% increase** over the default model (89.1%).

### Optimal Parameters

- n_estimators: 200 trees
- max_depth: 20 levels
- min_samples_leaf: 2
- min_samples_split: 5

# Results and Critical Insights

## Classification Performance by Stress Level

The model demonstrates robust, balanced performance across all three stress categories, with F1-scores—which balance precision and recall—remaining consistently high. This balanced performance is crucial for real-world application, ensuring that students at any stress level are identified with equal reliability.

**Overall Model Accuracy: 90.9%** – This exceeds the 85% clinical reliability threshold, making the model suitable for deployment in educational settings as a screening tool.


## Critical Finding: The Body Speaks First

The most powerful stress predictor is the physiological marker **Blood Pressure**, not traditional psychological indicators. This reveals that stress manifests physically before students may be psychologically aware or self-report elevated anxiety or depression levels.

### Actionable Strategy for Schools

This insight fundamentally shifts intervention strategy from purely psychological screening methods to incorporating physical health monitoring as a primary detection channel. Schools should consider routine physiological assessments—such as blood pressure checks during health screenings—as frontline indicators for mental health support needs, enabling earlier identification and intervention before psychological symptoms become severe.

# Research Foundation and Technical Framework

### Dataset Source

Ovi, M. S. I. (2024). *Student Stress Monitoring Dataset*. Kaggle. Comprehensive collection of 1,100 student records with 20 validated predictor variables across behavioral, physical, and environmental domains.

The dataset is from a **cross-sectional study**, meaning it **cannot establish causation**. The data is **self-reported**, introducing potential **response bias**.

### Methodological Foundation

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1). Seminal research establishing the theoretical foundation for ensemble learning methods and their application to complex classification problems.

### Technical Implementation

**Programming Environment:** Python with Jupyter Notebook for reproducible research

**Core Libraries:** Scikit-learn (model training), XGBoost (gradient boosting), Matplotlib (visualization)

**Version Control:** All code documented and version-controlled for transparency

## Data Processing

- Feature engineering
- Correlation analysis
- Class balancing verification

## Model Development

- Algorithm comparison
- Cross-validation
- Hyperparameter tuning

## Evaluation Metrics

- Accuracy and F1-scores
- Confusion matrices
- Feature importance

# Practical Deployment and Real-World Application

## From Model to Actionable Tool

The predictive model has been serialized and saved as `stress_predictor_model.pkl`, making it deployment-ready for integration into web-based assessment platforms or school management information systems.

The final assessment tool goes beyond simple classification (e.g., "High Stress") to provide personalized, explainable insights for each student. When a prediction is made, the system highlights specific risk factors contributing to the outcome, such as "elevated blood pressure and poor sleep quality," enabling targeted, individualized intervention strategies.

**Model Limitation:** The model cannot account for **sudden life events or crisis situations** , and it **does not account for temporal changes** in stress levels.


---

## Implementation Benefits for Educational Settings

### Rapid Screening

Process entire student cohorts efficiently during routine health assessments, identifying at-risk individuals within seconds rather than relying solely on lengthy psychological questionnaires.

### Personalized Interventions

Feature importance rankings enable counselors to design tailored support plans addressing each student's specific risk factors, whether physiological, behavioral, or environmental.

### Evidence-Based Decision Making

Provides quantifiable, defensible metrics for allocating mental health resources and demonstrating intervention effectiveness to school leadership and governing bodies.

### Proactive Prevention

Identifies students before crisis escalation, shifting from reactive crisis management to proactive mental health promotion and early intervention protocols.

# Future Directions and Model Enhancement

## Longitudinal Data Collection

Transition from single-timepoint predictions to **longitudinal tracking**, collecting measurements over time to enable prediction of stress progression trajectories and risk of academic burnout. This temporal dimension would allow schools to identify not just current stress levels but also concerning trends requiring preventive action. Conduct future research to **include demographic variables** (age, gender, culture) to identify at-risk populations.

## Enhanced Explainability with SHAP

Fully integrate **SHAP (SHapley Additive exPlanations)** values into the assessment interface. SHAP provides precise, granular mathematical breakdowns of how each input feature contributed to the final prediction, enhancing transparency, clinical trust, and stakeholder confidence in automated screening.

## Multi-School Validation Studies

Conduct external validation across diverse educational settings—varying by socioeconomic status, geographic location, and cultural context—to ensure model generalizability and identify any necessary calibration adjustments for different student populations.

## Mobile Assessment Platform

Develop smartphone-based self-assessment tools allowing students to voluntarily monitor their own stress indicators, fostering mental health literacy and self-advocacy whilst providing additional data streams for model refinement.

## Recommendations

Develop **sleep hygiene programs** and allocate resources for **students' basic needs** (housing, food security) as top priorities for intervention.

"By combining physiological monitoring with machine learning, we can transform student mental health support from reactive crisis response to proactive, data-informed wellbeing promotion."

# Thank You

I extend my sincere gratitude for your attention and engagement today.

Your insights and feedback are invaluable.