Student Stress Prediction

Predicting Student Stress for Actionable Interventions Using Machine Learning

This project aims to develop a robust machine learning model for **multi-class stress prediction** among students. By analyzing various stress factors, our goal is to identify key indicators and provide insights for effective interventions, thereby promoting academic success and mental well-being in educational settings.

Luis Pablo Aiello DATA ANALYST

Dataset Snapshot: Characteristics and Source

Understanding the Underlying Causes and Their Impact on Today's Students

This dataset investigates the root causes of stress among students, derived from a nationwide survey. It includes around 20 key features grouped under five scientifically identified categories:

🧠 Psychological Factors

- anxiety_level
- self_esteem
- mental_health_history
- depression

🌆 Environmental Factors

- noise_level
- living_conditions
- safety
- basic_needs

🤝 Social Factors

- social_support
- peer_pressure
- extracurricular_activities
- bullying

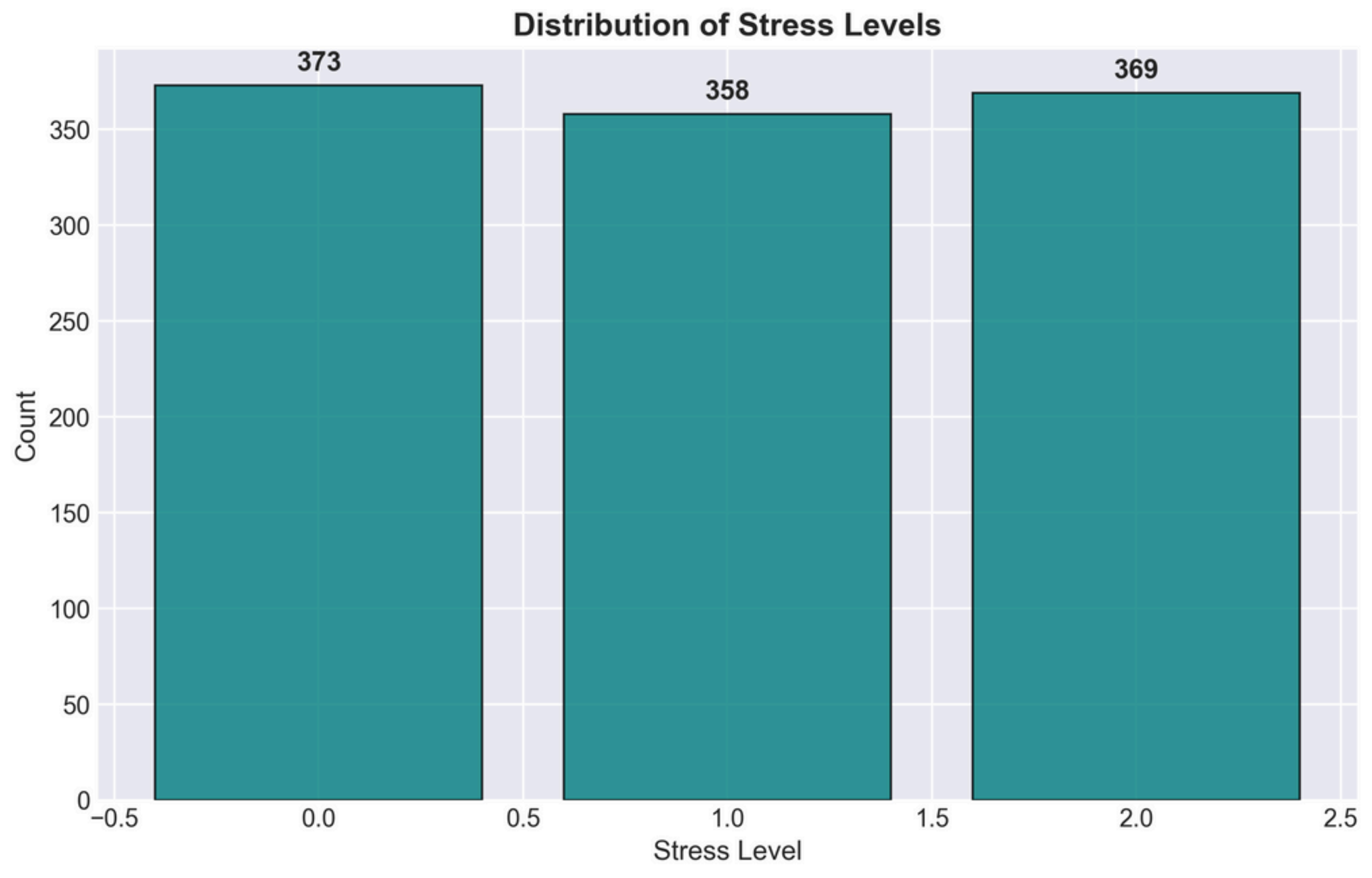
🏥 Physiological Factors

- headache
- blood_pressure
- sleep_quality
- breathing_problem

🎓 Academic Factors

- academic_performance
- study_load
- teacher_student_relationship
- future_career_concerns

Source: Kaggle — Student Stress Monitoring dataset



Data Overview

Rows: ~1,100

Features: ~21 (all numeric)

Missing values: none

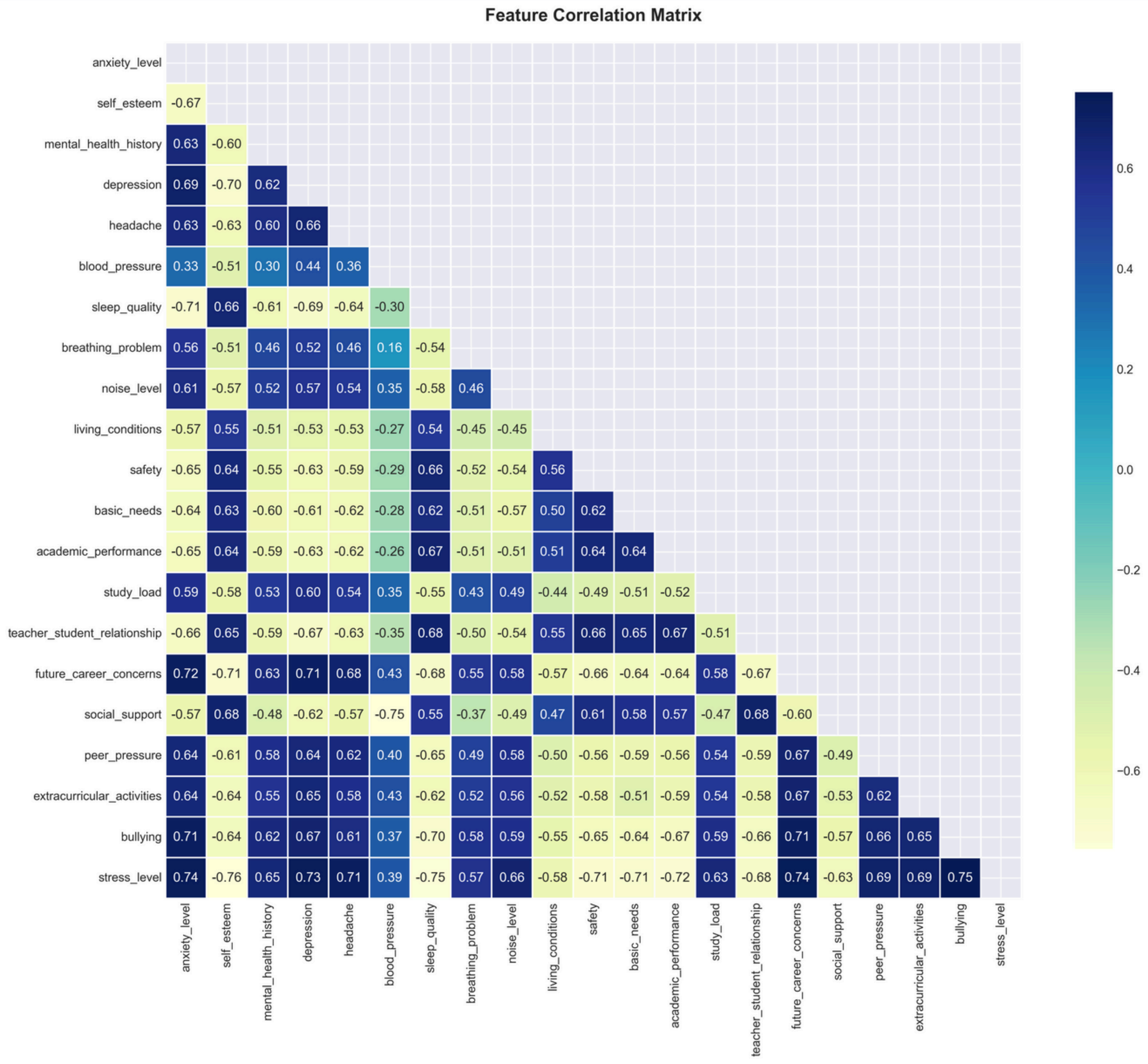
Duplicates: none

Target distribution (stress_level): {0: 373, 1: 358, 2: 369} → roughly balanced, so accuracy is a fair primary metric.

✅ Model-ready: numeric, balanced, clean

Exploratory Data Analysis

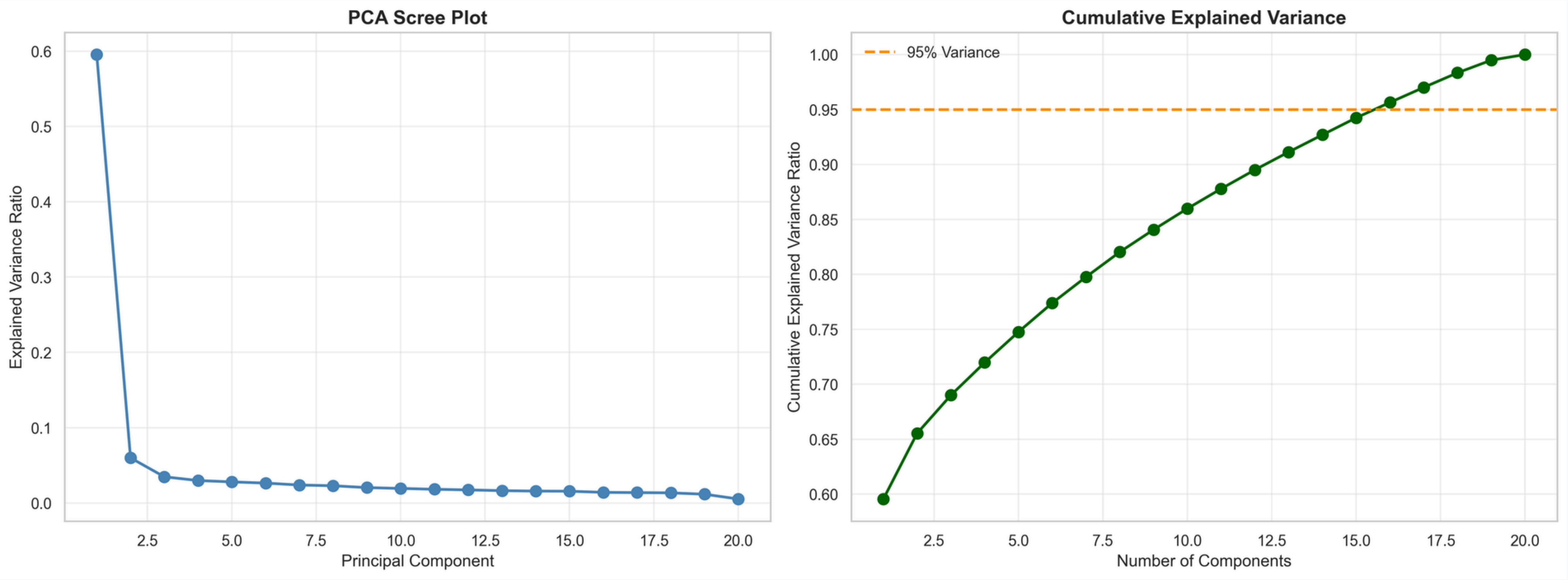
The analysis reveals key trends in student stress factors, highlighting correlations with **bullying**, **anxiety**, and protective elements like **sleep** and **self-esteem**, crucial for effective interventions.



High-correlation pairs (|r| > 0.70):

anxiety_level – future_career_concerns: **r = 0.717**
future_career_concerns – bullying: **r = 0.711**
anxiety_level – bullying: **r = 0.710**
depression – future_career_concerns: **r = 0.707**
anxiety_level – sleep_quality: **r = −0.710**
self_esteem – future_career_concerns: **r = −0.713**
blood_pressure – social_support: **r = −0.753**

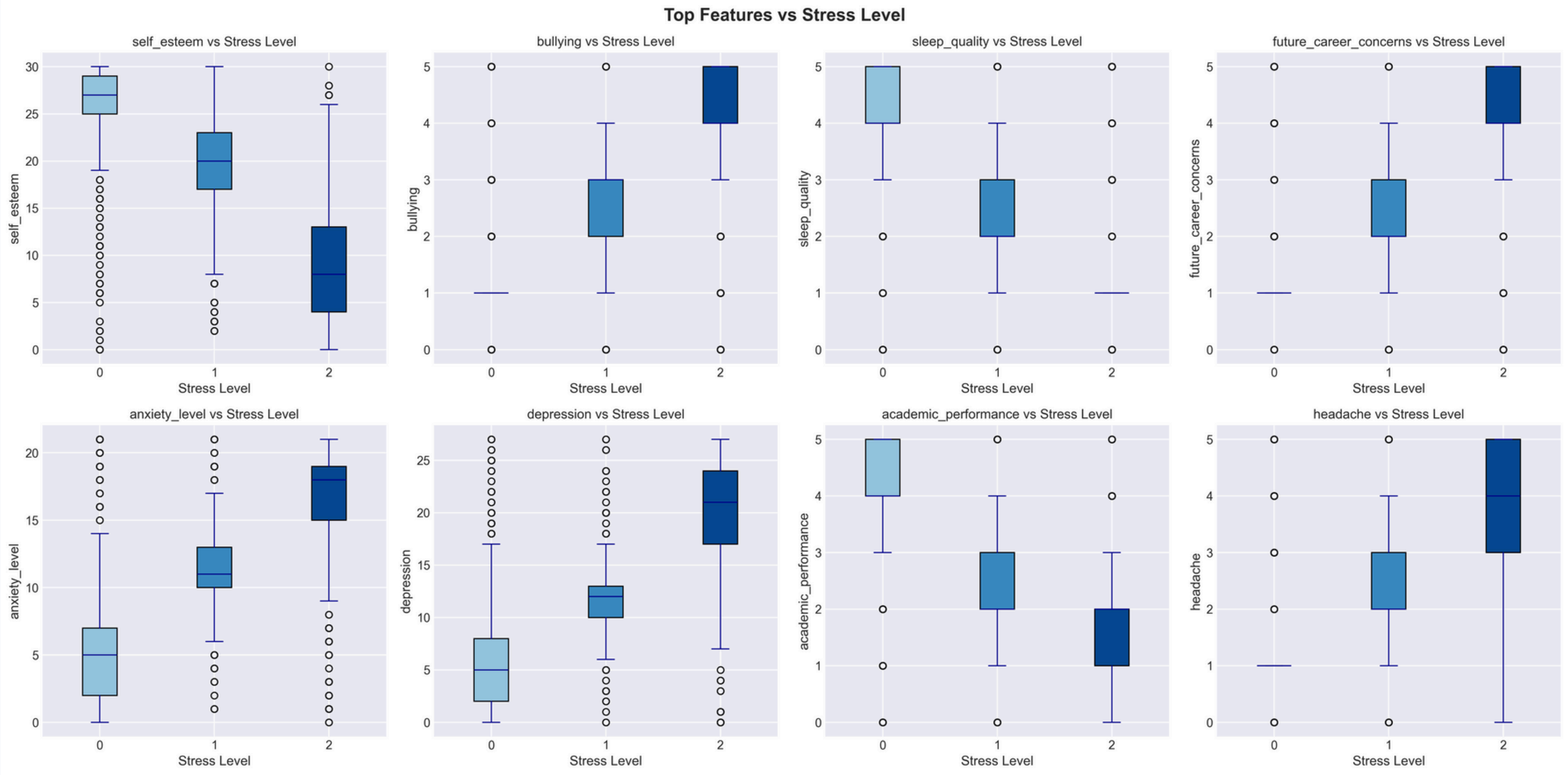
We drop: anxiety_level, future_career_concerns, social_support
We keep: bullying, depression, sleep_quality, self_esteem, blood_pressure

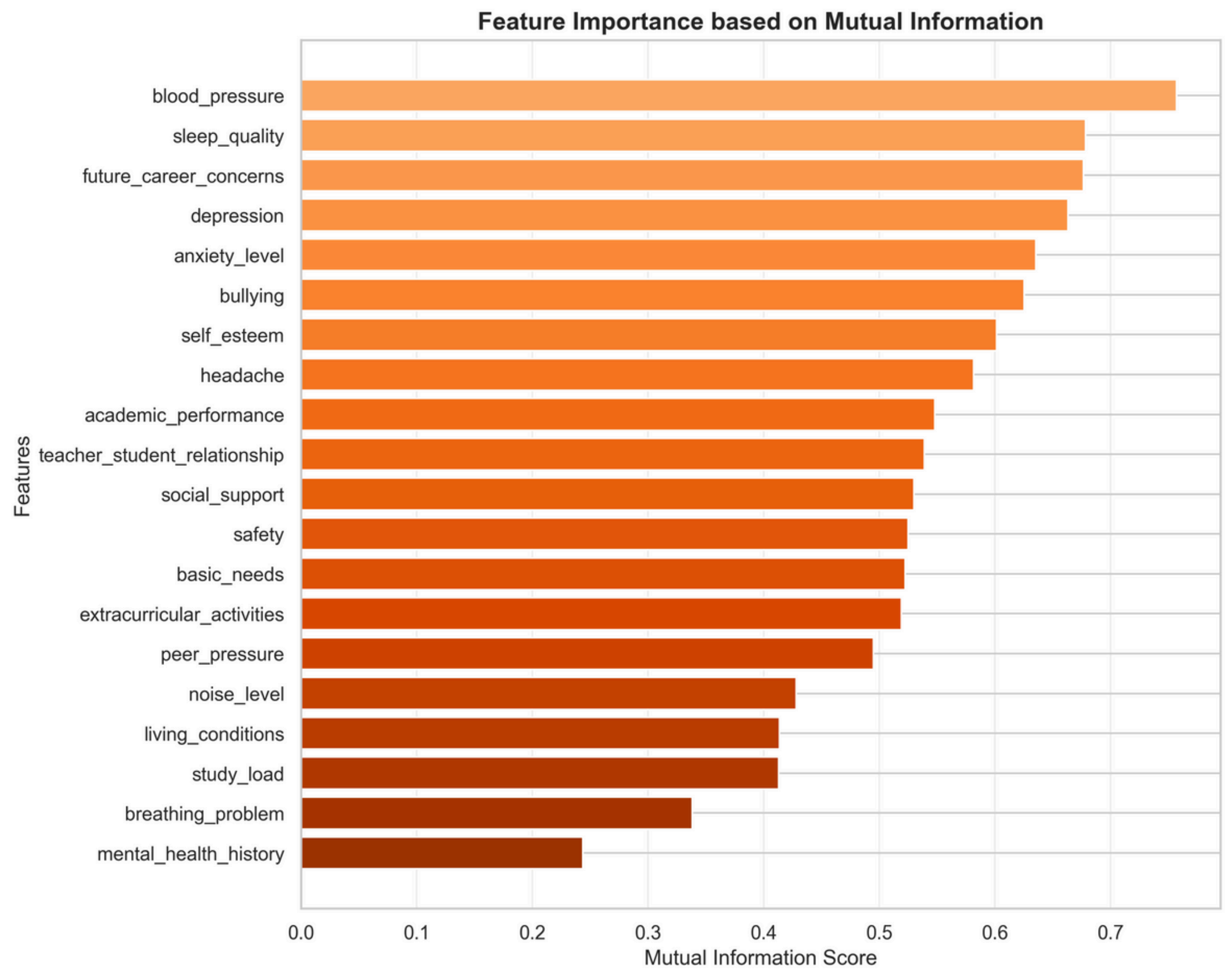


About 16 components explain **95%** of the variance; PCA can be used as an optional step for compact representations.

Stress Factors: Bullying, Anxiety, and Depression

Top MI features: blood_pressure, sleep_quality, future_career_concerns, depression, anxiety_level — these carry the most information about stress levels.





Key Stressors

Bullying, anxiety, and depression impact student well-being.

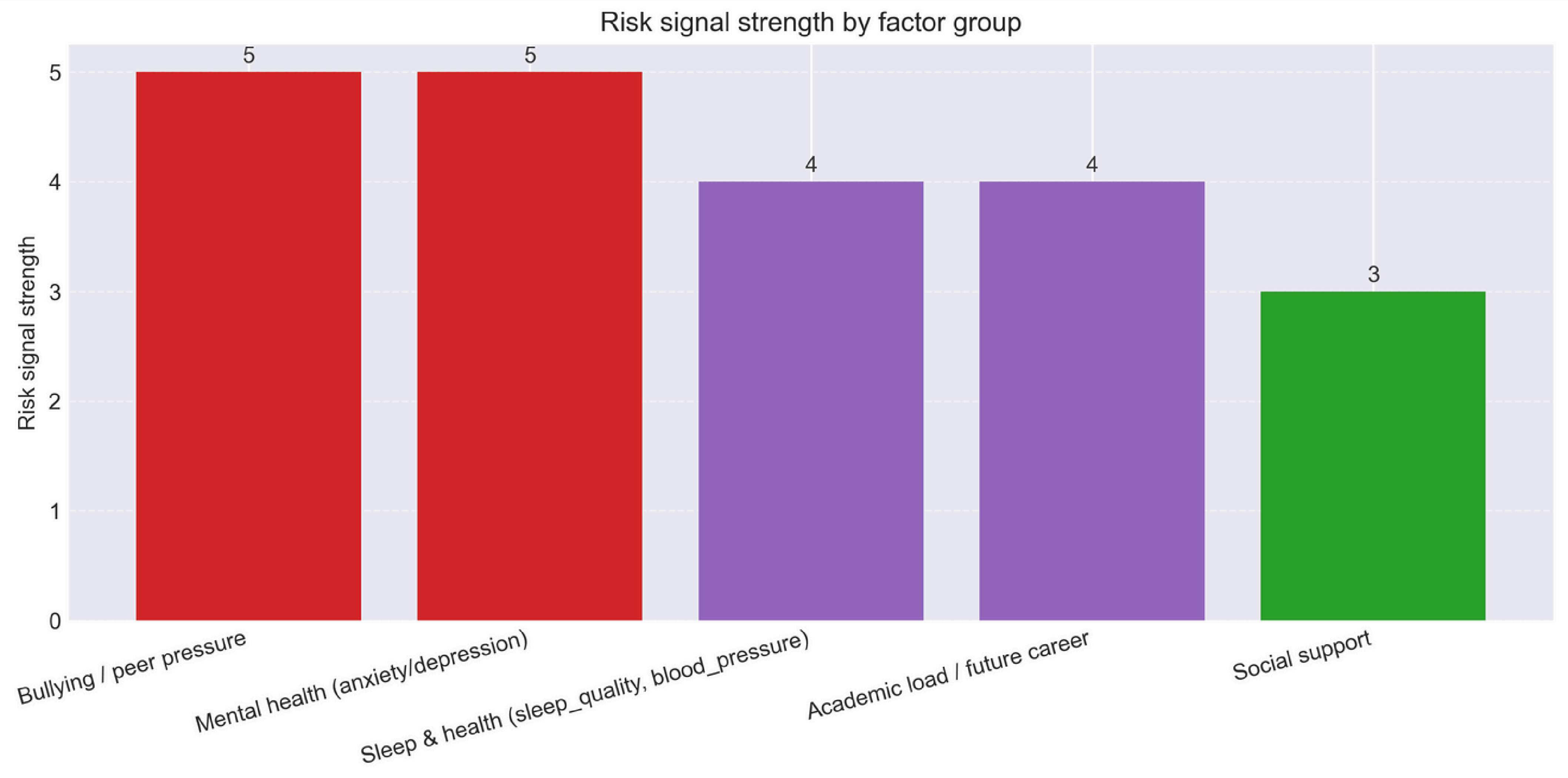
↑ **stress →** bullying, anxiety, depression, study_load, future_career_concerns, headache
↓ **stress →** sleep quality, self-esteem, academic_performance

Recommendations:

Turn Exploratory Data Analysis into Actions

What the school/program should do:

- **Reduce social stressors:** targeted anti-bullying / peer-pressure policies
- **Support mental health:** counseling, early screening for anxiety/depression
- **Protect recovery:** sleep & wellness / fatigue awareness
- **Lower academic** pressure: review study load, deadlines, exam clustering
- **Guide the future:** career mentoring to reduce uncertainty



Modeling Approach

Selecting Robust Models for Accurate Predictions

Target: **multi-class (0, 1, 2)**
Shared preprocessing → same for every model
Models: **KNN, Logistic Regression, Decision Tree**
Ensembles: **Random Forest, Gradient Boosting, AdaBoost**
Resampling runs: LogReg + **SMOTE / ROS / RUS**
Evaluation: **accuracy + macro precision/recall/F1**

In our project, we employed **leak-free pipelines** to ensure model integrity while exploring various algorithms, including **KNN**, **Logistic Regression**, and **Tree-based** models. By utilizing **ensemble techniques** and **resampling strategies**, we enhanced predictive performance. This approach not only improves accuracy but also provides a reliable framework for deploying effective solutions in real-world scenarios.

Model	Test Accuracy
Random Forest	**0.8955**
Gradient Boosting + hyperparams	**0.8955**
Logistic Regression + RUS	**0.8909**
Logistic Regression + SMOTE	**0.8909**
Logistic Regression	0.8864
Logistic Regression + ROS	0.8864
Random Forest + hyperparams	0.8818
AdaBoost + hyperparams	0.8818
AdaBoost	0.8773
Gradient Boosting	0.8591
Decision Tree	0.8545
KNN (k=17, CI-LCB)	0.8545

Results (Global Comparison)

Results — which models performed best?

Model	Accuracy	Macro Precision	Macro Recall	Macro F1	Notes
LogReg + SMOTE	0.8864	0.8863	0.8868	0.8863	Best recall for stressed classes (1, 2); slightly lower precision.
AdaBoost	0.8773	0.8809	0.8776	0.8774	Good overall but just below the top cluster.
Logistic Regression	0.8864	0.8888	0.8865	0.8873	Strong, easy to explain & deploy.
LogReg + ROS	0.8864	0.888	0.8865	0.8871	
LogReg + RUS	0.8909	0.8933	0.891	**0.8918**	Simpler; improves class-0 recall; very close to RF.
Random Forest	**0.8955**	0.896	0.8956	**0.8953**	Most diagonal confusion matrix; very balanced across 0 / 1 / 2.
Gradient Boosting (tuned)	**0.8955**	0.8598	0.8594	0.8596	Ties RF in accuracy after HP tuning; confirm macro balance.
Decision Tree	0.8545	0.8558	0.8551	0.8541	Lower, as expected; useful for teaching/baseline only.
KNN (k=17)	0.8545	0.8693	0.8552	0.856	Lower, as expected; useful for teaching/baseline only.

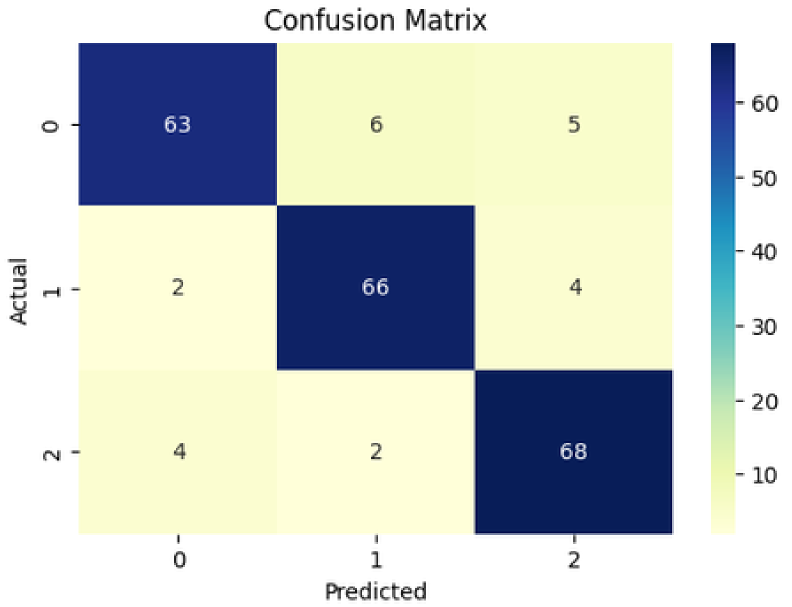
Per-Class Performance & Confusion Matrix

Model	Per-class Precision			Per-class Recall			Per-class F1		
	Class 0	Class 1	Class 2	Class 0	Class 1	Class 2	Class 0	Class 1	Class 2
Random Forest	**0.913**	0.8919	0.8831	0.8514	**0.9167**	**0.9189**	0.8811	**0.9041**	**0.9007**
LogReg + RUS	0.8442	**0.9559**	0.88	0.8784	**0.9028**	0.8919	0.8609	**0.9286**	0.8859
LogReg + SMOTE	0.8421	0.8649	0.8533	0.8649	**0.9167**	0.8919	0.8533	**0.9362**	0.8859
Logistic Regression	0.8421	**0.9559**	0.8684	0.8649	**0.9028**	0.8919	0.8533	**0.9286**	0.880
Gradient Boosting	0.8267	**0.9014**	0.8514	0.8378	0.8889	0.8514	0.8322	0.8951	0.8514

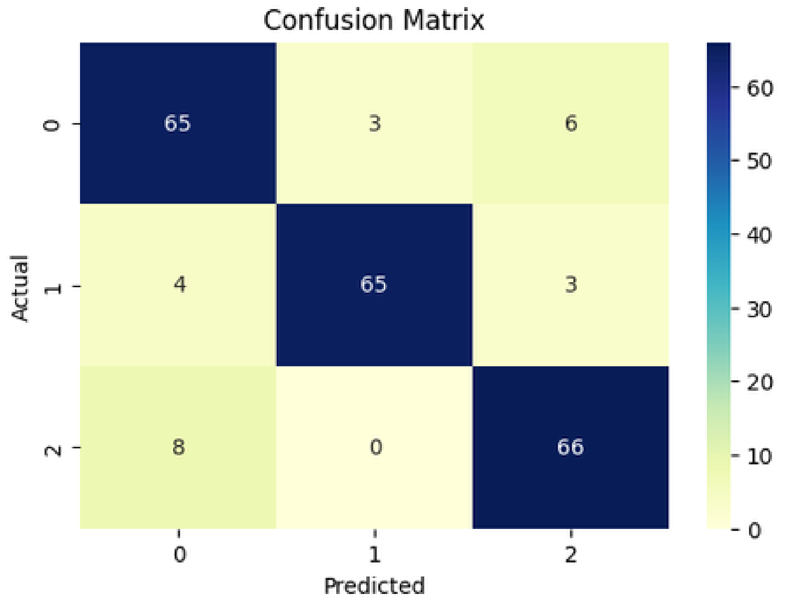
Table: Random Forest — Per-Class Metrics

Class	Precision	Recall	F1-score
0 (low)	0.9130	0.8514	**0.8811**
1 (medium)	0.8919	0.9167	**0.9041**
2 (high)	0.8831	0.9189	**0.9007**

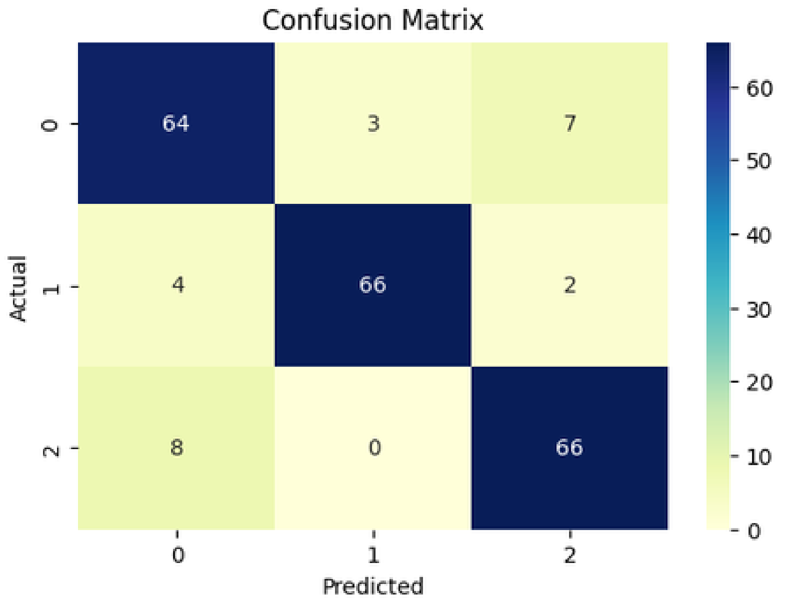
=== Random Forest ===



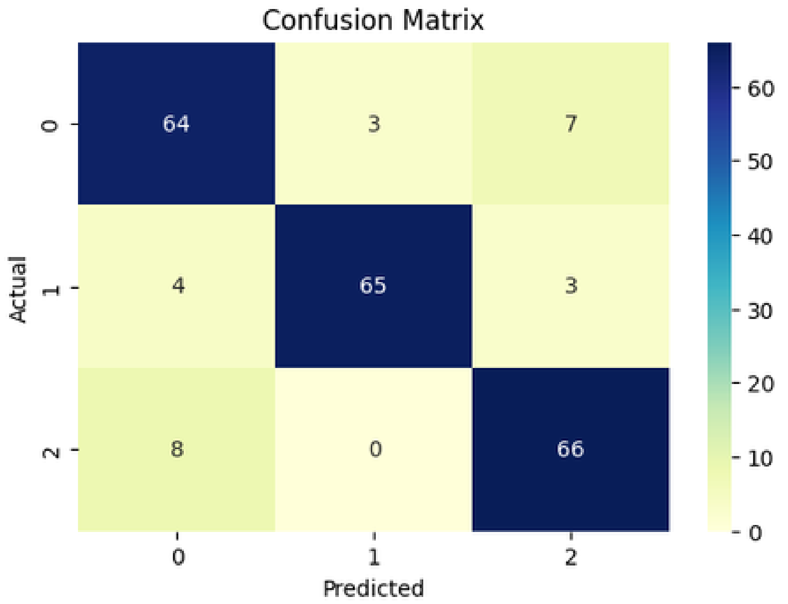
=== LogReg + RUS ===



=== LogReg + SMOTE ===



=== Logistic Regression===



Model Selection & Top Features

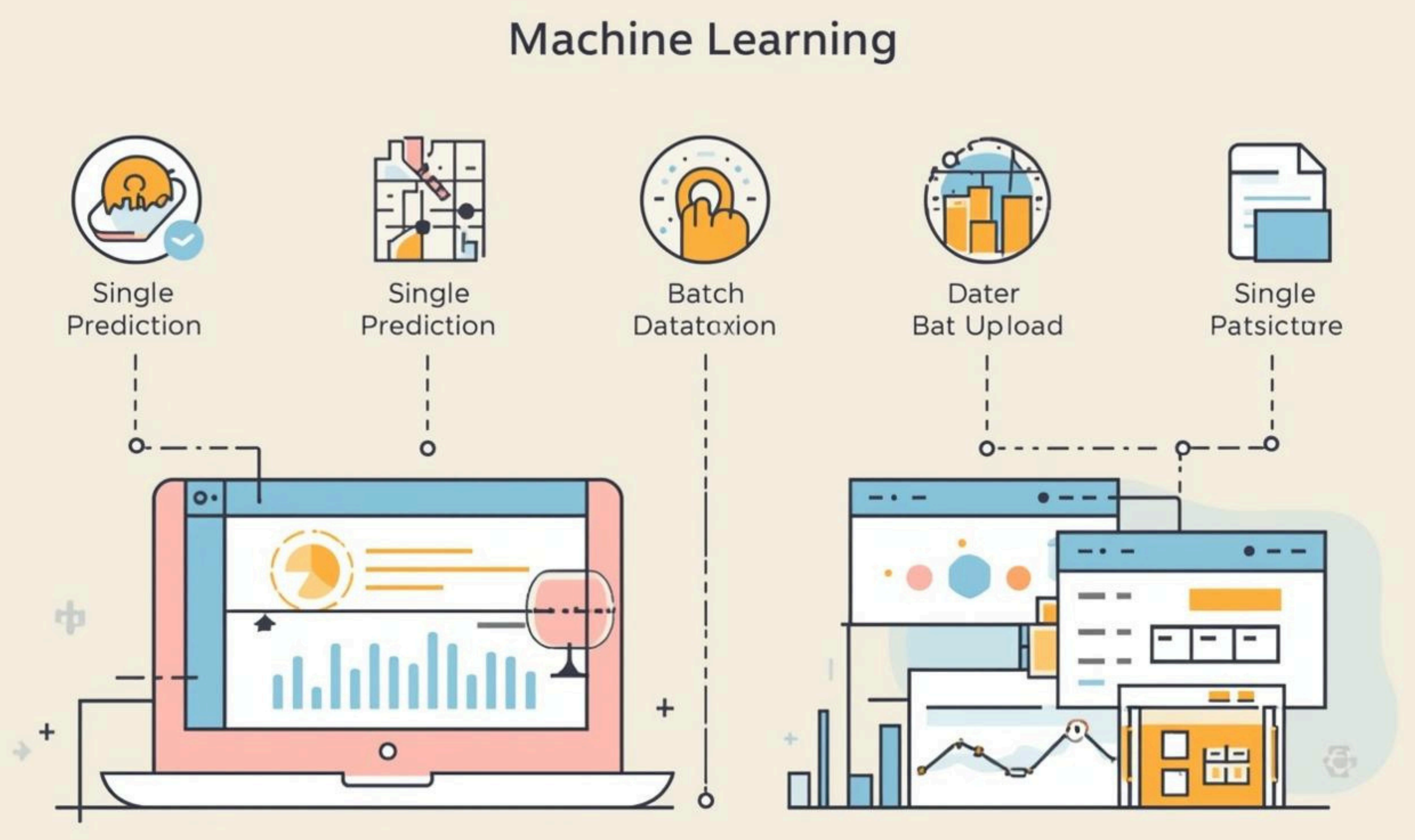
Recommended Models for Deployment

Model	Why keep it	When to use it	Export status / notes
Random Forest	Best overall balance; top accuracy (**0.8955**); diagonal confusion	General-purpose, demo, default model	Exported as .pkl; uses full feature schema
Gradient Boosting (tuned)	Ties RF after HP tuning; competitive on same schema	When you can afford slightly heavier model	Exported if tuned pipeline was fitted
LogReg + SMOTE	Same accuracy band (**0.8909**) but **better recall** on stressed 1/2	When "catch stressed students" is the KPI	Exported as .pkl; good for explanations
LogReg + RUS	Very close (0.8909) but **simpler** and interpretable	When stakeholders want coefficients and simplicity	Exported as .pkl; maps to same feature list
Logistic Regression	Strong baseline (**0.8864**); trivial to explain	Fallback / educational / fast scoring	Exported as .pkl

Top signals: blood_pressure, sleep_quality, social_support, anxiety_level, depression, self_esteem, academic_performance, study_load, future_career_concerns, bullying. That tells us the model is learning the same story we saw in EDA.

Demo Overview: Predictions Made Easy

Feature	Effect on stress	Explanation (short)
self_esteem	**↑ → ↓ stress**	more confidence, less distress
mental_health_history	**↑ → ↑ stress**	past issues → higher risk
depression	**↑ → ↑ stress**	direct psychological load
headache	**↑ → ↑ stress**	physical symptom of stress
blood_pressure	**↑ → ↑ stress**	physiological activation
sleep_quality	**↑ → ↓ stress**	better sleep → better coping
breathing_problem	**↑ → ↑ stress**	health discomfort → more stress
noise_level	**↑ → ↑ stress**	worse environment → more stress Kaggle
living_conditions	**↑ → ↓ stress**	better housing → safer → calmer Kaggle
safety	**↑ → ↓ stress**	feeling safe lowers stress
basic_needs	**↑ → ↑ stress**	in this dataset "needs not met" → more stress Kaggle
academic_performance	**↑ → ↓ stress**	performing well → less pressure Medium
study_load	**↑ → ↑ stress**	heavier workload → more stress Medium
teacher_student_relationship	**↑ → ↓ stress**	support from teachers helps
peer_pressure	**↑ → ↑ stress**	social pressure is a stressor ✅ (fixed)
extracurricular_activities	**↑ → ↓ stress**	engagement can buffer stress ✅ (fixed) Medium
bullying	**↑ → ↑ stress**	direct social stressor



Predictions Simplified

Easily make predictions with individual or bulk data inputs.

Link: → Student Stress Prediction

Stress Signals

Understanding Key Indicators and Performance Metrics

The analysis revealed crucial **stress signals** affecting students, such as **bullying**, **anxiety**, and **depression**. Our models effectively predicted these indicators, demonstrating high accuracy and relevance. By addressing these challenges, we can implement actionable interventions that provide targeted support, ultimately fostering a healthier academic environment. Continuous monitoring and adaptation will enhance our response strategies.

Student Stress

Feel free to reach out with any questions!

WEBSITE

www.reallygreatsite.com

EMAIL

hello@reallygreatsite.com

PHONE

123-456-7890