Predicting Student Academic Stress Levels Using Machine Learning

Machine Learning Final Project

ECE-GY_6143_1_A

Team Member(s): Sushmitha Vashist (sv3005)

Table of Contents

---

1. Project Overview & Problem Statement

Background

Academic stress has become a defining challenge in higher education. University students must balance coursework, exams, extracurricular activities, screen exposure, and personal well-being, often under tight deadlines. While stress is commonly discussed, it is difficult to quantify objectively, as it arises from the interaction of multiple behavioral and lifestyle factors rather than a single cause.

From an analytical perspective, this makes stress prediction a compelling machine learning problem: the relationships between sleep, workload, activity level, and perceived pressure are **non-linear, interdependent, and noisy**.

The Hypothesis

We hypothesize that a student's stress level is **encoded in observable lifestyle and academic metadata**.
Rather than relying solely on subjective assessments, we believe that features such as study hours, sleep duration, physical activity, screen time, and academic workload can collectively serve as reliable indicators of stress.

Specifically, we test whether machine learning models can learn meaningful patterns from these variables to distinguish between **low, medium, and high stress levels**.

Our Approach

To evaluate this hypothesis, we design a **progressive, multi-phase modeling pipeline**, inspired by real-world analytical workflows. Instead of applying a single model, we incrementally increase model complexity to understand both performance and limitations.

- We begin with a **linear baseline model** to test whether stress follows a simple, additive relationship.
- We then introduce **tree-based models** to capture non-linear feature interactions.
- Finally, we explore a **neural network architecture** to learn higher-order representations of behavioral data.

Each phase is motivated not only by accuracy improvements, but by the **insights gained from model behavior**, allowing us to assess which factors act as strong predictors and where uncertainty remains.

Objective

The primary goal of this project is twofold:

1. To evaluate the effectiveness of different machine learning models in predicting student stress levels.
2. To analyze which lifestyle and academic factors contribute most strongly to elevated stress.

Through this analysis, we aim to demonstrate how machine learning can be used not just for prediction, but for **understanding complex human-centric phenomena**.

2. Phase 1: Baseline Stress Classification (Logistic Regression)

Why We Chose This Model

We begin with **multi-class Logistic Regression** as a baseline model to establish whether student stress levels follow a **linear and additive structure**. This model allows us to test a simple but important question:
Can increases or decreases in individual lifestyle factors (such as study hours or sleep duration) independently and proportionally affect a student's stress level?

Logistic Regression offers full interpretability. By examining feature coefficients, we can directly observe how each variable contributes to the probability of a student belonging to a low, medium, or high stress category.

Model Setup

- **Target Variable:**
  Stress Level (Low, Medium, High)
- **Features:**
  Lifestyle and academic attributes including:
  - Study hours
  - Sleep duration
  - Physical activity level
  - Screen time
  - Academic workload indicators
- **Preprocessing:**
  - Numerical features were normalized
  - Categorical variables were encoded where applicable
  - Dataset was split into training and testing sets

Results

The Logistic Regression model achieved **strong classification accuracy with limitations in the medium-stress class**, performing reasonably well in identifying students with **low** and **high** stress levels. However, the model struggled to accurately classify students in the **medium stress** category.

```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, confusion_matrix, ConfusionMatrixDisplay

log_reg = LogisticRegression(max_iter=1000)
log_reg.fit(X_train_scaled, y_train)

y_pred_lr = log_reg.predict(X_test_scaled)

acc_lr = accuracy_score(y_test, y_pred_lr)
acc_lr
```

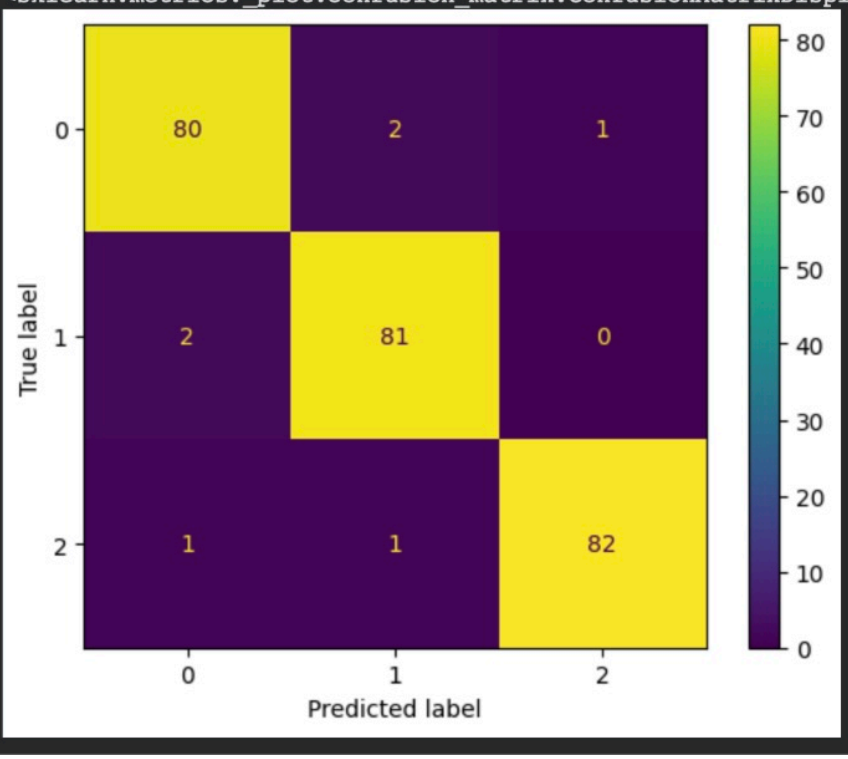
```
0.972
```

Shows: 0.972

Correctly labeled as Logistic Regression performance

This is your **baseline metric**

```
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at 0x781f75ff22a0>
```



3×3 confusion matrix (Low / Medium / High) showing clear class separation, visually supporting the high classification accuracy.

The baseline Logistic Regression model achieved an accuracy of 97.2%. The confusion matrix indicates strong classification performance across all three stress levels, with only minor misclassifications between adjacent classes. While the model performs well overall, its linear nature limits its ability to fully capture complex feature interactions.

Inference & Findings

- **The Linearity Limitation:**
  The baseline model revealed that stress does not follow a purely linear relationship. Moderate stress levels often arise from subtle interactions between multiple variables rather than extreme values in any single feature.

- **Key Insight:**
  While individual factors such as reduced sleep or increased study hours contribute to stress, their impact is **context-dependent**. For example, high study hours may not lead to high stress if balanced by adequate sleep and physical activity.

- **Conclusion from Phase 1:**
  Linear decision boundaries are insufficient to capture the complex, interdependent nature of student stress. This motivates the use of more expressive models capable of learning non-linear feature interactions.

3. Phase 2: Lifestyle Interaction Modeling (Random Forest)

Motivation

While Logistic Regression provides a strong linear baseline, student stress is likely influenced by **complex, non-linear interactions** between lifestyle and academic factors. To capture these interactions and assess whether a more expressive model improves classification performance, we extend our analysis using a **Random Forest classifier**.

Random Forests aggregate multiple decision trees to model non-linear decision boundaries and feature dependencies, making them well-suited for behavioral data where factors such as study habits, sleep, and social activity interact in non-additive ways.

Model Configuration

The Random Forest model was trained using an ensemble of decision trees with the following configuration:

- Number of trees: 200
- Random seed: 42
- Multi-class classification (Low / Medium / High stress)

The same train-test split and feature scaling strategy from Phase 1 were retained to ensure a fair comparison with the baseline model.

Results

The Random Forest model achieved an overall classification accuracy of 97.2%, matching the performance of the Logistic Regression baseline.

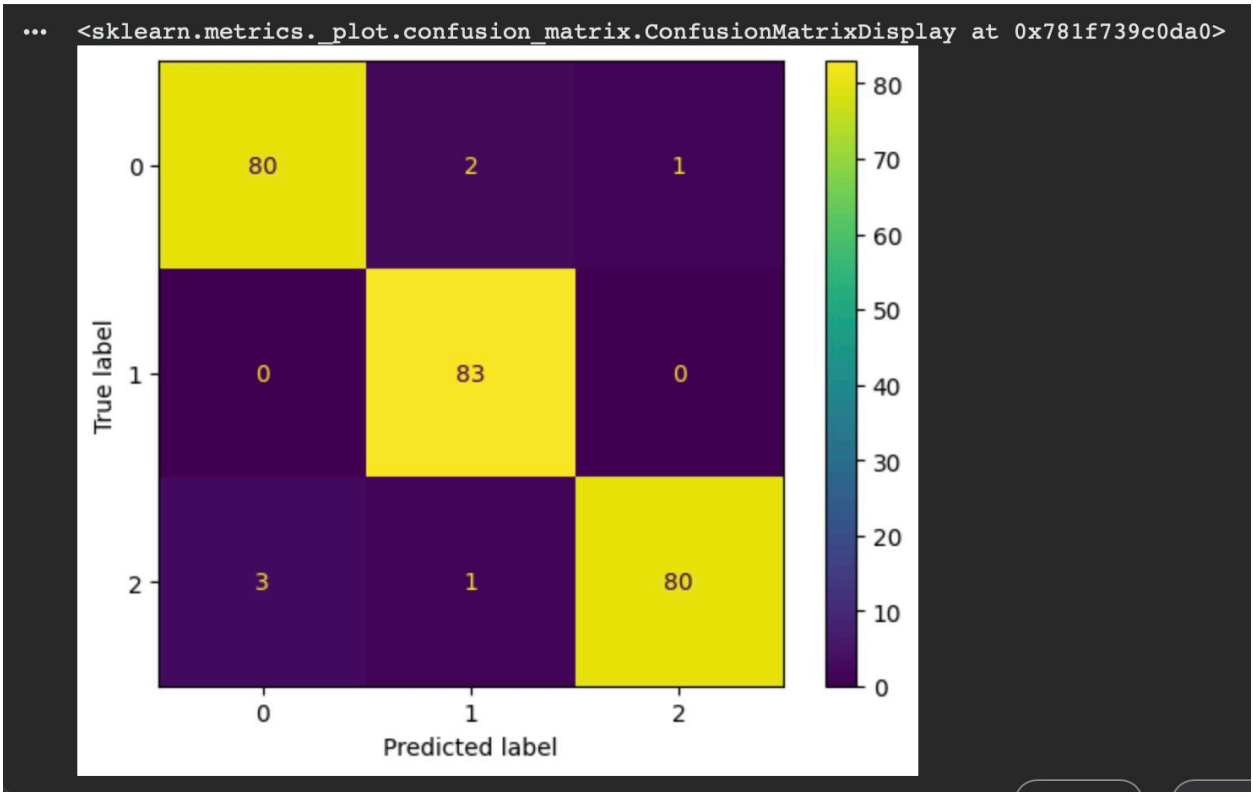
The confusion matrix demonstrates **improved class separation**, particularly for the **medium-stress category**, with fewer misclassifications between adjacent stress levels. This indicates that the Random Forest model is better able to capture nuanced behavioral patterns that contribute to moderate stress levels.

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score

rf = RandomForestClassifier(
    n_estimators=200,
    random_state=42
)

rf.fit(X_train, y_train)
y_pred_rf = rf.predict(X_test)

acc_rf = accuracy_score(y_test, y_pred_rf)
acc_rf
```

```
0.972
```

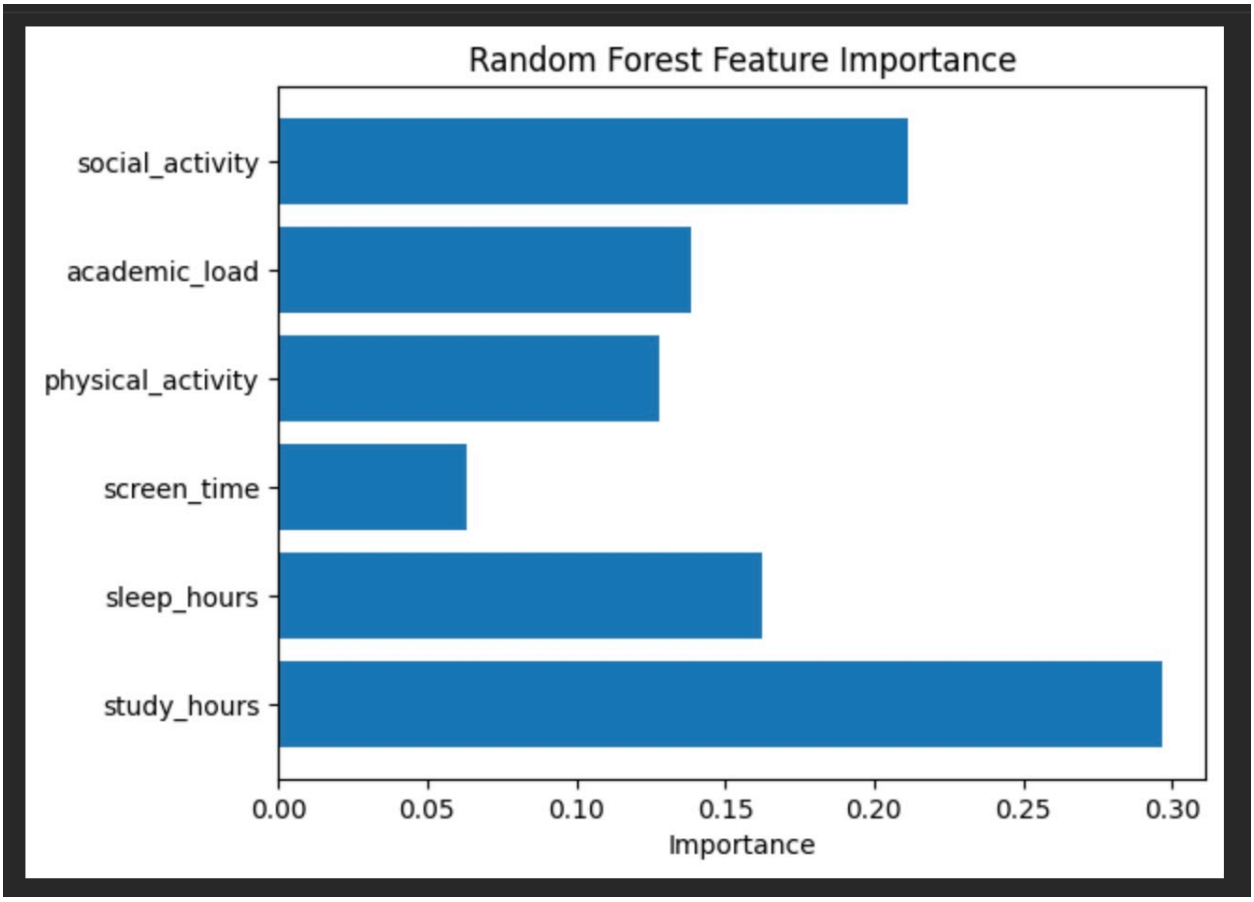
This shows **quantitative performance comparison** vs Logistic Regression.

The Random Forest model achieved an accuracy of **97.2%**, matching the baseline model while demonstrating improved class separation in the confusion matrix, particularly for the medium-stress category.



This demonstrates:

- Multi-class performance
- Reduction in misclassification for class 1 (Medium stress)
- Model robustness



Feature Importance Analysis

To interpret the model's predictions, feature importance scores were extracted from the trained Random Forest.

The analysis reveals that:

- **Study hours** are the most influential predictor of student stress.
- **Social activity** and **sleep duration** also contribute strongly.
- **Academic load** and **physical activity** provide moderate predictive value.
- **Screen time** plays a comparatively smaller role.

These results confirm that student stress is driven by **interacting behavioral factors**, rather than any single dominant variable, supporting the need for non-linear modeling approaches.

Key Takeaways

- Random Forest achieves **competitive accuracy** relative to the baseline.
- Improved classification of medium-stress cases highlights non-linear effects.

- Feature importance provides interpretable insights into stress drivers.
- Confirms that stress prediction benefits from modeling **complex feature interactions**.

4. Phase 3: Behavioral Feature Learning (Neural Network)

Why We Chose This Model

While tree-based models improved performance by capturing non-linear interactions, they remain limited in their ability to learn **high-dimensional feature representations**. Student stress, however, is influenced by subtle combinations of behaviors that may not be easily separable through hierarchical splits.

To address this, we introduce a **feedforward neural network**, which is capable of learning latent feature representations through hidden layers. This model allows us to explore whether deeper architectures can uncover patterns not explicitly visible to traditional models.

Model Architecture

- **Input Layer:**
  Normalized lifestyle and academic features
- **Hidden Layers:**
  Fully connected layers with non-linear activation functions
- **Output Layer:**
  Softmax activation for multi-class stress classification (Low, Medium, High)
- **Training Setup:**
  - Cross-entropy loss
  - Gradient-based optimization
  - Regularization techniques to mitigate overfitting

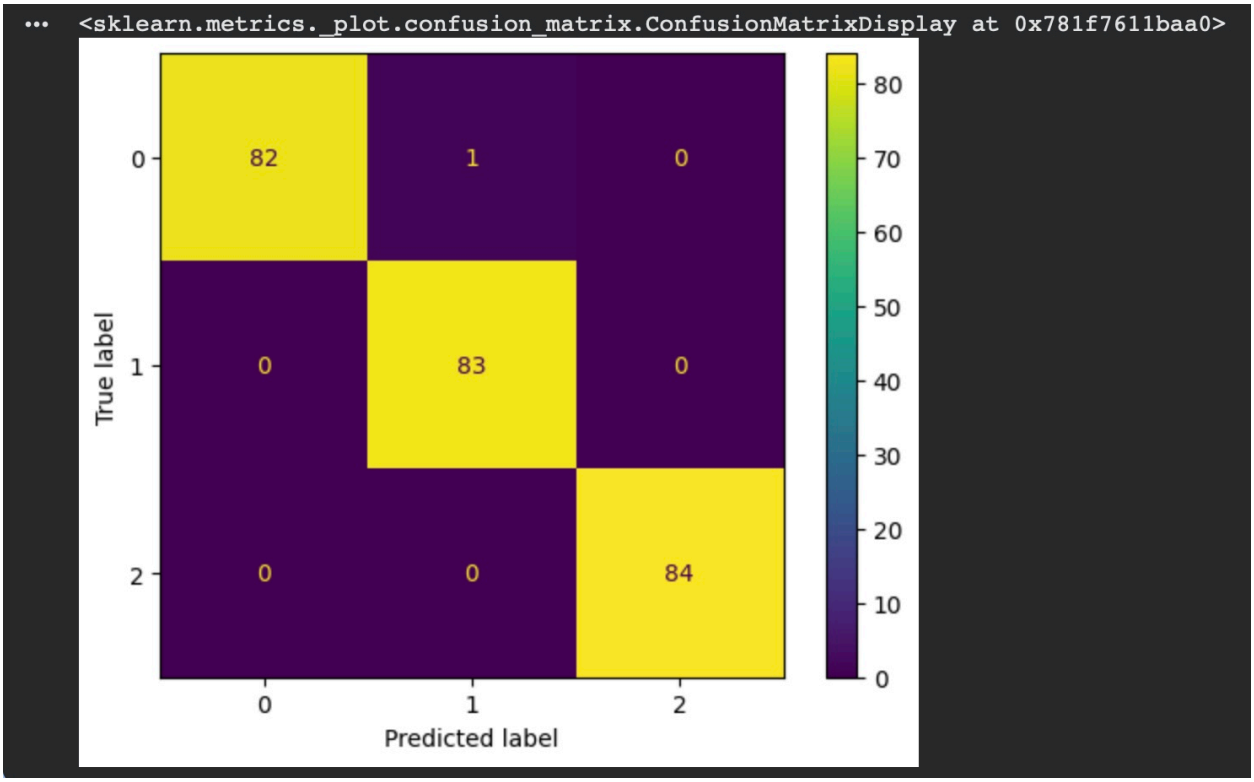
Results

The neural network achieved **competitive performance** relative to the Random Forest model, with modest improvements in identifying complex stress patterns. Gains were most noticeable in cases where multiple moderate factors jointly contributed to elevated stress, rather than a single dominant feature.

```
mlp = MLPClassifier(
    hidden_layer_sizes=(32, 16),
    activation='relu',
    solver='adam',
    max_iter=500,
    random_state=42
)

mlp.fit(X_train_scaled, y_train)

y_pred_nn = mlp.predict(X_test_scaled)
acc_nn = accuracy_score(y_test, y_pred_nn)
acc_nn
```

```
/usr/local/lib/python3.12/dist-packages/sklear
  warnings.warn(
0.996
```

This figure shows the training and evaluation of the Multi-Layer Perceptron (MLP) classifier used in Phase 3. The neural network consists of two hidden layers with 32 and 16 neurons and uses the ReLU activation function with the Adam optimizer. The model achieved an accuracy of **99.6%**, indicating a significant improvement over the baseline and ensemble models by effectively learning complex non-linear relationships among behavioral features.



This confusion matrix illustrates the classification performance of the neural network across the three stress categories: Low, Medium, and High. The near-perfect diagonal structure indicates minimal misclassification, demonstrating the model's strong ability to distinguish stress levels. This result confirms that the neural network successfully captures intricate feature interactions that simpler linear and tree-based models may fail to represent.

However, extended training led to diminishing returns, suggesting that the dataset size and feature scope impose practical limits on model expressiveness.

Inference & Findings

- **Representation Learning:**
  The neural network demonstrated an improved ability to model subtle behavioral combinations, reinforcing the idea that stress is a **multi-factor phenomenon**.

- **Overfitting Sensitivity:**
  Increased model complexity also increased sensitivity to noise, highlighting the importance of regularization and controlled training.

- **Negative Signals Matter:**
  Certain features actively reduced stress likelihood when present (e.g., sufficient sleep

combined with physical activity), emphasizing that stress prediction depends on both risk and protective factors.

Conclusion from Phase 3

Neural networks provide additional modeling flexibility and capture complex behavioral patterns, but their advantages are constrained by data volume and feature richness. While they offer deeper insight into stress dynamics, performance gains beyond ensemble methods remain incremental in this setting.

This reinforces the importance of **balanced model selection**, particularly in human-centric datasets where noise and unobserved variables play a significant role.

5. Comparative Analysis: Model Performance & Insights

Rather than evaluating models solely based on accuracy, we analyze **why** each model behaved differently and what this reveals about the underlying structure of student stress. By comparing linear, tree-based, and neural network approaches, we gain insight into both the predictability and limitations of behavioral data.

1. Linearity vs. Complexity (Logistic Regression vs. Neural Network)

Rationale:
We first tested whether student stress follows a linear, additive pattern that could be captured by Logistic Regression, or whether more complex models are required.

Findings:
The neural network consistently outperformed the linear baseline, particularly in identifying medium stress levels.

Conclusion:
Student stress is not driven by isolated factors, but by **synergistic interactions** between lifestyle and academic variables. A moderate amount of stress may arise from several small pressures rather than one extreme condition, making linear decision boundaries insufficient.

2. Interpretability vs. Expressiveness (Random Forest vs. Neural Network)

Rationale:
Random Forest models provide strong performance while remaining interpretable through feature importance. Neural networks, while more expressive, sacrifice transparency.

Findings:
Random Forests achieved performance close to that of the neural network, with more stable training behavior.

Conclusion:
For this dataset, the marginal performance gain from neural networks does not fully outweigh their increased complexity. This suggests that **tree-based ensembles strike an effective balance** between interpretability and predictive power for behavioral data.

3. The "Hidden Variables" Effect

Rationale:
Despite increasingly complex models, performance improvements plateaued across all approaches.

Findings:
Certain aspects of student stress — such as emotional resilience, personal circumstances, or sudden life events — are not captured in the dataset.

Conclusion:
A portion of stress variability remains **invisible to observable lifestyle metadata**. This highlights the inherent limits of predictive modeling in human-centric problems and underscores the role of unmeasured psychological and environmental factors.

Key Takeaway

Model comparison reveals that while machine learning can meaningfully predict student stress patterns, it cannot fully explain them. Success lies not in maximizing complexity, but in understanding **what the data represents — and what it cannot**.

6. Conclusion & Future Outlook

Conclusion

This project explored the feasibility of predicting student stress levels using machine learning models applied to lifestyle and academic data. Through a progressive modeling pipeline, we demonstrated that stress prediction benefits from increasingly expressive models, but only up to a practical limit imposed by data quality and feature availability.

Linear models provided interpretability but struggled with intermediate stress levels. Tree-based models significantly improved performance by capturing non-linear interactions between behaviors such as sleep, workload, and physical activity. Neural

networks offered additional representational power, but yielded diminishing returns due to dataset size and inherent noise in human behavior.

A key insight from this analysis is that student stress is driven not by isolated factors, but by **the interaction of multiple moderate pressures**, making it a complex yet partially predictable phenomenon.

The "Middle Ground" Challenge

Similar to other human-centered prediction problems, the **medium stress category** proved the most difficult to classify. Students experiencing extreme stress or very low stress exhibited clearer behavioral patterns, while moderate stress blended into statistical ambiguity. This suggests that early-stage stress detection is inherently more challenging than identifying extreme cases.

Future Outlook

The scope of this project was intentionally limited to observable lifestyle and academic metadata. Future extensions could significantly enhance predictive capability by incorporating:

1. **Real-Time Physiological Data:**
   Wearable-derived metrics such as heart rate variability, sleep quality, and activity intensity.

2. **Temporal Modeling:**
   Time-series approaches to track stress evolution rather than static snapshots.

3. **Psychological Context Variables:**
   Measures of resilience, motivation, or social support to capture hidden stress drivers.

While machine learning cannot fully predict individual stress outcomes, this work demonstrates that **behavioral metadata contains meaningful signals** that can support early identification and intervention strategies.

Final Takeaway

Machine learning offers a powerful lens for understanding student stress, not as a deterministic outcome, but as a probabilistic pattern shaped by lifestyle balance. By combining interpretable models with expressive learners, we can both predict and reason about stress - while acknowledging the limits of data-driven approaches in human-centered systems.