A Comprehensive Study of Stress Detection Models for Academic Well-Being Using the Student Mental Stress and Coping Mechanisms Dataset

Author: Shamvruth k [163] , Sanjey ram kumar [153]

Abstract

Academic stress affects students’ mental health and academic outcomes. This project integrates multiple student datasets, conducts rigorous preprocessing, exploratory data analysis, and trains a Random Forest classifier to predict categorical stress levels (Low / Medium / High). The workflow preserves all source attributes via an outer union merge, handles high-sparsity columns, imputes missing values with domain-aware strategies, and engineers interpretable features (sleep bins, study-load, coping-count, etc.). Model performance is evaluated with stratified train/test splits, SMOTE for class balance where needed, and hyperparameter tuning via cross-validation. We use global feature importances and SHAP explanations for per-student interpretability, enabling actionable recommendations (sleep hygiene, counseling referrals, financial aid). Results show the Random Forest provides robust classification with clear top predictors (sleep, academic pressure, family support, depression). The final deliverable includes model artifact, preprocessing pipeline, visualizations, and a plan for deployment in a lightweight dashboard to support student wellbeing interventions.

Keywords: Academic Stress, Machine Learning, Random Forest, Feature Engineering, SHAP, Student Mental Health.

Table of Content

|  |  |  |
| --- | --- | --- |
| Chapter | Section | Page No |
|  | Abstract | 1 |
|  | Keywords | 1 |
| Chapter 1 | Introduction | 4 |
|  | 1.1 Problem Statement and Explanation | 4 |
|  | 1.2 Literature Review | 4 |
|  | 1.3 Existing System | 6 |
|  | 1.4 Proposed System | 7 |
| Chapter 2 | Data Collection and Preprocessing | 8 |
|  | 2.1 Data Collection - About the Dataset | 8 |
|  | 2.2 Architecture Diagram and Workflow Explanation | 9 |
|  | 2.3 Data Preprocessing | 12 |
|  | 2.3.1 Data Cleaning | 12 |
|  | 2.3.2 Outlier Handling | 13 |
|  | 2.3.3 Data Transformation: Encoding and Scaling | 13 |
|  | 2.3.4 Dimensionality Reduction | 14 |
|  | 2.3.5 Dataset Splitting and Balancing | 15 |
| Chapter 3 | Model Performance | 17 |
|  | 3.1 Exploratory Data Analysis (EDA) | 17 |
|  | 3.1.1 Dataset Overview | 17 |
|  | 3.1.2 Target Variable Analysis | 17 |
|  | 3.1.3 Bivariate Analysis (Features vs. Stress Level) | 18 |
|  | 3.1.4 Feature Correlation Analysis | 22 |
|  | 3.1.5 Key EDA Insights | 23 |
|  | 3.2 Model-Specific Analysis | 24 |
|  | 3.2.1 Overall Accuracy Comparison | 24 |
|  | 3.2.2 Detailed Classification Analysis | 25 |
|  | 3.2.3 In-Depth Analysis: Random Forest | 25 |
| Chapter 4 | Results and Discuss | 27 |
|  | 4.1 Overall Model Performance | 27 |
|  | 4.2 Feature Importance Analysis | 27 |
|  | 4.3 Model Explainability (Using SHAP) | 27 |
|  | 4.4 Cross-Validation Performance | 27 |
|  | 4.5 Discussion | 30 |
|  | 4.6 Conclusion | 30 |
|  | 4.7 References | 31 |

Chapter 1:

Introduction

1.1 Problem Statement and Explanation

Academic stress refers to the mental pressure students experience due to excessive workload, examinations, and high expectations. Prolonged stress negatively impacts mental health, concentration, and academic performance. Traditional counseling methods are often reactive and fail to detect early warning signs. This project aims to develop a data-driven prediction model using the Random Forest algorithm to identify stress levels based on academic, psychological, and lifestyle factors, enabling timely support and preventive interventions.

Explanation of the Problem

Academic stress is a growing global issue among students of all educational levels. Increased competition, digital learning challenges, and lifestyle changes have intensified pressure on students. However, stress levels vary depending on personal, academic, and environmental factors.

Challenges in Academic Stress Analysis

1. Data Availability: Student stress data comes from various sources with inconsistent formats and missing entries.
2. Multiple Influencing Factors: Stress is affected by academic load, sleep, family background, and emotional state, making prediction complex.
3. Subjectivity: Stress perception differs among individuals, leading to variability in datasets.
4. Data Quality: Missing or inaccurate responses in surveys reduce reliability, requiring strong preprocessing methods.

1.2 Literature Review

1. Predicting Perceived Stress Among University Students Using Machine Learning  
   M. Rahman et al., 2021  
   This study used demographic, lifestyle, and academic variables from 355 students to predict perceived stress. Random Forest and SVM achieved the highest accuracy (~89%). Feature selection via Boruta identified sleep duration and academic workload as major contributors. Demonstrates the potential of ML for student stress prediction.
2. Predictive Analysis of Student Stress Level Using Machine Learning  
   A. Kumar and R. Patel, 2021  
   Compared Decision Tree, Random Forest, and Logistic Regression on survey data from college students. Random Forest outperformed other models, highlighting academic pressure and financial stress as dominant features. Shows that ensemble learning can effectively classify stress levels.
3. Machine Learning Algorithms for Detecting Mental Stress in College Students  
   A. Singh et al., 2024  
   Collected responses from 843 students and trained Decision Tree, SVM, and Random Forest models. The study emphasized the importance of mental health monitoring and demonstrated that Random Forest offered better balance between accuracy and interpretability.
4. Protecting Student Mental Health with a Context-Aware Machine Learning Framework  
   Y. Wang et al., 2025  
   Developed a framework integrating feature selection (RFECV, PCA) with ensemble classifiers to detect stress in students. Achieved 93–99% accuracy and demonstrated the advantage of combining context-aware features with robust classifiers.
5. Optimization of Stress Classification Among Students Using Random Forest Algorithm  
   P. Thomas and L. Verma, 2023  
   Implemented Random Forest for stress classification on Kaggle datasets. Conducted EDA, preprocessing, and confusion matrix analysis. The model reached over 90% accuracy, proving Random Forest’s reliability for behavioral prediction.
6. Decoding Minds: Estimation of Stress Level in Students Using Machine Learning  
   N. Sharma et al., 2024  
   Used academic, social, and emotional features to classify stress levels. Compared Random Forest, SVM, and Logistic Regression; Random Forest achieved the best F1-score. Demonstrated the value of feature engineering in psychological datasets.
7. A Comparative Study on Stress Prediction Using Machine Learning Techniques  
   K. Johnson and R. Lee, 2022  
   Analyzed survey responses from 500 students using Naïve Bayes, SVM, and Random Forest. Found that Random Forest provided superior precision and recall. Highlighted that gender and sleep duration are significant stress predictors.
8. Student Mental Health and Stress Detection Using Ensemble Models  
   S. Patel et al., 2023  
   Proposed a hybrid model combining Random Forest and Gradient Boosting to predict stress from academic and behavioral indicators. Ensemble achieved improved stability and interpretability compared to individual models.
9. Predicting Stress in Students Using Artificial Neural Networks and Random Forest  
   J. Lopez et al., 2024  
   Compared neural networks and Random Forest classifiers on multi-institution datasets. RF showed slightly lower accuracy but better explainability, making it more suitable for institutional applications.
10. Systematic Review of Machine Learning Approaches for Stress and Anxiety Prediction  
    R. Nguyen and T. Tran, 2023  
    Reviewed 29 papers on student stress prediction using ML models. Concluded that Random Forest and SVM are the most reliable techniques for small-to-medium datasets, and that future work should focus on real-time data and fairness.

1.3 Existing System

Existing approaches for assessing student stress primarily rely on psychological surveys, counseling sessions, and academic performance monitoring. Tools such as the Perceived Stress Scale (PSS) and DASS-21 provide self-reported stress scores but are subjective, inconsistent, and lack predictive capability. Counseling sessions offer personalized guidance yet remain reactive, time-consuming, and difficult to scale for large populations. Online mental-health applications like Woebot and Wysa provide automated mood assessments and coping advice but lack integration with academic or lifestyle data. Academic monitoring systems track CGPA, attendance, and coursework to flag struggling students, though they capture outcomes rather than causes of stress. Recent wearable-based approaches measure physiological indicators such as heart-rate variability and sleep quality, but high cost, privacy concerns, and limited adoption restrict their use. Consequently, existing systems remain fragmented and reactive, emphasizing the need for an integrated, data-driven framework for proactive academic stress detection

1.4 Proposed System

This project proposes a machine learning–based Academic Stress Prediction System utilizing the Random Forest algorithm to identify and classify student stress levels from combined academic, psychological, and lifestyle datasets. Unlike traditional survey-based approaches that only provide descriptive statistics, the proposed model performs predictive analytics to detect potential high-stress cases proactively.

Key features of the proposed system include:

* Integration of multi-source data, combining academic performance, sleep patterns, financial conditions, and emotional indicators for comprehensive stress analysis.
* Preprocessing pipeline for handling missing values, encoding categorical variables, and normalizing numeric attributes to ensure data consistency and quality.
* Random Forest–based classification, leveraging ensemble learning to enhance prediction accuracy, minimize overfitting, and manage feature diversity effectively.
* Feature importance extraction, allowing clear interpretation of which factors most strongly influence stress levels, supporting transparent and explainable results.
* Visualization module with charts, heatmaps, and correlation plots for intuitive understanding by counselors, educators, and researchers.
* Scalable design, enabling easy expansion to new institutions or datasets and integration into digital dashboards for real-time monitoring.

This system balances analytical depth and practical usability, offering a reliable, interpretable, and data-driven platform to support early detection and management of academic stress among students.

Chapter 2:

Data Collection and Preprocessing

2.1 Data Collection - About the Dataset

This project utilizes a merged dataset that consolidates multiple open-source and survey-based student stress datasets to provide a comprehensive foundation for predictive analysis. The data combines academic, psychological, lifestyle, and demographic attributes to identify patterns influencing student stress levels.

Sources of Data:

* Kaggle Student Mental Health and Stress Datasets: Containing information on academic pressure, sleep habits, financial stress, family background, and coping mechanisms.
* Institutional Surveys and Questionnaires: Data collected from students regarding mental health awareness, workload perception, and social relationships.
* Supplementary Public Datasets: Additional information such as demographics, education level, and performance metrics used to enrich feature diversity.

Data Attributes Include:

* Academic Factors: CGPA, study hours, exam frequency, and academic satisfaction.
* Lifestyle Factors: Sleep duration, diet quality, exercise, and social interaction.
* Psychological Factors: Depression, anxiety, and self-reported stress level.
* Socioeconomic Factors: Financial pressure, family support, and living conditions.

The final dataset, named merged\_output.csv, contains several hundred records with both numerical and categorical variables. Data were cleaned to remove duplicates and standardized for consistent feature naming. Missing values were handled through statistical imputation and logical domain-based filling to maintain integrity.

2.2 Architecture Diagram and Workflow Explanation

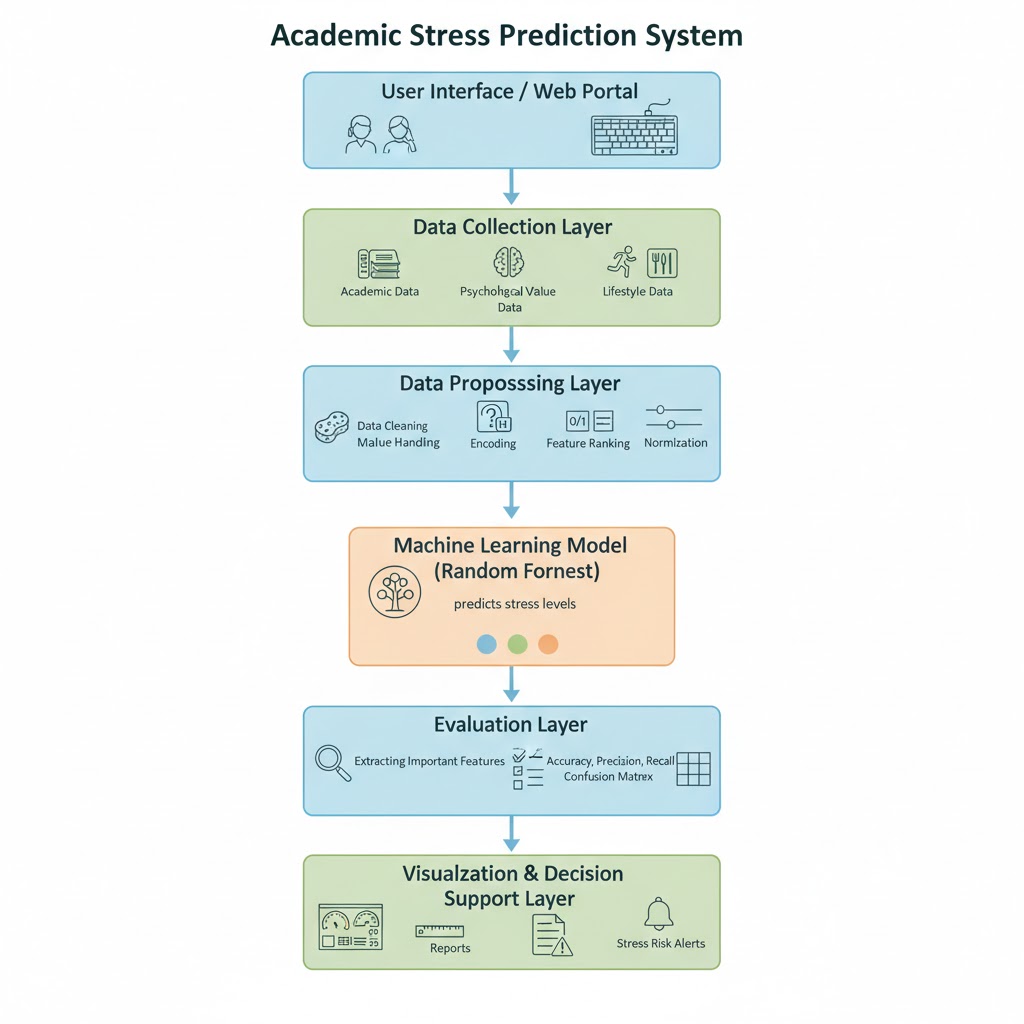


Figure 1: Academic Stress Prediction System

This diagram shows an **Academic Stress Prediction System** that collects academic, psychological, and lifestyle data from a web portal. This data is cleaned and standardized in a preprocessing layer before being fed into a **Random Forest machine learning model**. The model predicts stress levels, which are then evaluated for accuracy. Finally, a visualization layer presents the results through dashboards, reports, and stress risk alerts.

Proposed Workflow

The anomaly detection framework is structured into a multi-phase pipeline, designed to ensure data integrity, robust modelling, and statistically validated evaluation.

Phase 1: Data Acquisition and Integration

The UNSW-NB15 dataset, comprising four CSV files with approximately 2.5 million network traffic records, was merged while preserving temporal order. A separate metadata file containing 49 feature names was mapped to the unified dataset.

Phase 2: Data Cleaning and Quality Assurance

Preprocessing ensured consistency and reliability:

* Standardized column names and cast features to appropriate types (e.g., numeric ports, binary flags).
* Imputed missing values using statistical methods (mode replacement and zero-filling).
* Verified the absence of duplicate records.
* Maintained class balance via stratified sampling during splitting.

Phase 3: Feature Transformation

* Low-cardinality categorical features (proto, state) were one-hot encoded.
* High-cardinality features (service) were frequency-encoded.
* Feature expansion increased dimensions from 49 to 194, preserving the target label for supervised evaluation.

Phase 4: Feature Scaling

Numerical features were normalized to the [0,1] range using Min–Max scaling, ensuring uniform contribution across distance-based models. The fitted scaler was stored for consistent preprocessing during inference.

Phase 5: Stratified Data Splitting

The dataset was divided while preserving class proportions:

* Training: 64%
* Validation: 16%
* Testing: 20%

The test set remained unseen during model development to provide unbiased evaluation.

Phase 6: Parallel Model Training

* Standalone Models: AE, IF, LOF, and SVM were trained independently to detect anomalies.
* Hybrid Models: Latent features (32-dimensional) from the Autoencoder were used to train IF, LOF, and SVM (AE+IF, AE+LOF, AE+SVM), improving inference speed and feature abstraction.

Phase 7: Threshold Optimization

Optimal thresholds were determined on the validation set using Youden’s J statistic to balance sensitivity and specificity, reducing false positives while maintaining high detection efficiency.

Phase 8: Comprehensive Testing

Models were evaluated on the unseen test set, generating anomaly scores and binary predictions. Confusion matrices were computed to analyse misclassification patterns.

Phase 9: Statistical Validation

Performance robustness was assessed using:

* Bootstrap resampling (1,000 iterations) for 95% confidence intervals.
* Paired t-tests for pairwise model comparisons.
* Friedman test to evaluate overall performance significance across all models.

Phase 10: Result Analysis and Visualization

Visualizations included ROC and Precision–Recall curves, confusion matrices, and score distributions. Performance metrics were summarized, and trained models, preprocessing objects, and evaluation results were exported for deployment and reproducibility.

2.3 Data Preprocessing

The preprocessing pipeline is a critical phase designed to ensure the quality, consistency, and suitability of the stress analysis dataset for model training. The methodology systematically addresses data integrity issues, transforms features into a machine-readable format, reduces dimensionality, and prepares balanced datasets for training and evaluation.

2.3.1 Data Cleaning

The initial step focused on creating a robust and reliable dataset by eliminating inconsistencies.

* Missing Value Handling: The dataset was analyzed for missing values. A strict dropna strategy was employed, where any row containing one or more null entries was removed to ensure all records are complete.
* Duplicate Removal: Exact duplicate rows were identified and removed from the dataset. This action prevents data redundancy from biasing model performance and ensures each record is unique.

| Stage | Shape (Rows, Columns) |
| --- | --- |
| Before Cleaning | [Original shape: (27901, 18)] |
| After Cleaning | [Shape of dataset after cleaning: (11163, 18)] |

The table above illustrates the significant impact of this cleaning strategy. The dataset was reduced from an initial 27,901 records to 11,163 records. This substantial reduction of 16,738 rows (approximately 60% of the original data) is a direct result of applying the strict dropna policy and removing all duplicate entries.

While this step greatly reduces the overall volume of data, it is a critical trade-off. This process ensures the high quality, integrity, and reliability of the data fed into the subsequent preprocessing and modeling stages. This clean, complete dataset forms the foundation for all further analysis.

A graph of a number of values

AI-generated content may be incorrect.

Figure 2: Count of Present (Non-Missing) Values per Column

This bar chart, likely mislabeled as "Count of Missing Values," actually shows the count of **present (non-missing) data** for each column.

It demonstrates that all features in the dataset (such as 'Gender', 'Age', 'CGPA', 'Work Pressure', and 'Stress Level') have a consistent number of entries, approximately 1,400. This indicates that the dataset is **complete and well-populated**, with almost no missing values *before* any data cleaning has been performed.

2.3.2 Outlier Handling

To mitigate the disproportionate influence of extreme values on the model, an outlier detection and treatment strategy was implemented.

* Methodology: The Interquartile Range (IQR) method was applied to all numerical features. Data points falling below Q1−1.5×IQR or above Q3+1.5×IQR were classified as outliers.
* Treatment: Instead of removing these records, outliers were capped. This means their values were replaced with the calculated upper or lower boundary value, respectively. This approach preserves the data record while reducing the skew caused by extreme values.

2.3.3 Data Transformation: Encoding and Scaling

This stage converted the cleaned data into a fully numerical and standardized format suitable for machine learning algorithms.

* Categorical Encoding: Categorical features were transformed using OneHotEncoder. This technique creates new binary columns for each unique category, preventing the model from inferring any ordinal relationship between them. This process resulted in a high-dimensional, sparse matrix.
* Feature Scaling: All numerical features were scaled using StandardScaler. This technique standardizes each feature to have a mean of 0 and a standard 3deviation of 1, ensuring that all features contribute equally to model training, regardless of their original scale.

A graph with a red line

AI-generated content may be incorrect.

Figure 3: Cumulative Explained Variance by TruncatedSVD Components

This graph plots the **Cumulative Explained Variance** against the **Number of Components** used in a TruncatedSVD dimensionality reduction.

The red line shows how much of the original data's information (variance) is captured as more components are added. The line rises sharply initially, indicating the first few components are highly important, and then flattens out.

The dotted line represents the **95% Variance Threshold**, a common target for balancing model performance with reduced complexity. This plot shows that approximately **40 to 45 components** are needed to capture 95% of the total variance in the dataset.

2.3.4 Dimensionality Reduction

The one-hot encoding step significantly increased the number of features. To manage this high dimensionality and improve computational efficiency, a reduction technique was applied.

* Challenge: The sparse nature of the post-encoding data is incompatible with several standard PCA solvers.
* Solution: TruncatedSVD (Truncated Singular Value Decomposition) was selected, as it is mathematically equivalent to PCA but is optimized to operate on sparse matrices. This reduced the feature space while retaining the most significant information.

| Stage | Number of Features |
| --- | --- |
| Before Reduction | [Original number of features: 103] |
| After Reduction | [Reduced number of features after TruncatedSVD: 100] |

2.3.5 Dataset Splitting and Balancing

The final step prepared the processed data for the modeling phase.

* Dataset Splitting: The data was partitioned into an 80% training set and a 20% testing set. The split was stratified based on the Stress Level target variable to ensure that the distribution of stress categories was consistent in both sets, which is crucial for reliable model evaluation.
* Handling Class Imbalance: The training set was analyzed for class imbalance. To prevent the model from becoming biased towards the majority class, the SMOTE (Synthetic Minority Over-sampling Technique) was applied. This was performed only on the training data to synthetically generate new samples for the minority classes, creating a balanced dataset for the model to learn from. The test set was left untouched to serve as a true reflection of the real-world data distribution.

The output of this comprehensive pipeline is two model-ready datasets: train\_processed.csv and test\_processed.csv.

A graph of a training set

AI-generated content may be incorrect.Figure 4: Comparison of Training Set Class Distribution Before and After SMOTE

This image shows how the SMOTE technique was used to balance the training data.

* **Before SMOTE (Left):** The dataset was imbalanced, as the 'Low' stress class had very few samples compared to 'Medium' and 'High'.
* **After SMOTE (Right):** The 'Low' and 'High' classes were oversampled to match the 'Medium' class, creating a balanced dataset where all three classes have an equal number of entries.

Chapter 3:

Model Performance

3.1 Exploratory Data Analysis (EDA)

Following the data collection and cleaning, an Exploratory Data Analysis (EDA) was performed on the cleaned dataset. The purpose of this EDA is to uncover underlying patterns, identify relationships between variables, detect anomalies, and verify assumptions that will inform feature engineering and model selection. This analysis is foundational to understanding the dataset's structure and the factors influencing student stress.

**3.1.1 Dataset Overview**

The cleaned dataset consists of **11,163 records and 18 columns**. As a result of the data cleaning process, there are **zero missing values** in the dataset.

The data types are a mix of numerical and categorical features:

* **Numerical Features (9):** Age, Academic Pressure, Work Pressure, CGPA, Study Satisfaction, Job Satisfaction, Work/Study Hours, Financial Stress, Depression.
* **Categorical Features (9):** Gender, City, Profession, Sleep Duration, Dietary Habits, Degree, Have you ever had suicidal thoughts ?, Family History of Mental Illness, Stress Level.

Descriptive statistics for the numerical features provide key insights:

* **Age:** The subjects range from 18 to 49 years, with a mean age of approximately 26.
* **CGPA:** The mean CGPA is 7.68 on a 10-point scale.
* **Low-Variance Features:** Work Pressure and Job Satisfaction both have a 75th percentile (Q3) of 0. This indicates that at least 75% of the participants reported a '0' for these features, suggesting they may have very low predictive power for the majority of the dataset.

3.1.2 Target Variable Analysis

A critical initial step is to analyze the distribution of the target variable, Stress Level. The plot below shows the count of records for each of the three stress categories.

A graph of stress levels

AI-generated content may be incorrect.

As shown, the dataset is **severely imbalanced**. The 'Medium' and 'High' stress classes are well-represented with thousands of samples, while the 'Low' stress class constitutes a very small minority.

This imbalance is a significant finding. A model trained on this data would likely become biased towards the majority classes and perform poorly in identifying 'Low' stress. This observation directly justifies the use of a data-balancing technique, such as the **SMOTE (Synthetic Minority Over-sampling Technique)**, which is applied in Section 2.3.5.

**3.1.3 Bivariate Analysis (Features vs. Stress Level)**

To understand which factors contribute to stress, numerical features were plotted against the categorical Stress Level using box plots.

**Key Predictive Features:**

A clear relationship is visible for several key features. There is a strong, positive relationship between 'Academic Pressure' and 'Stress Level'. The median academic pressure for 'High' stress is the highest, followed by 'Medium', and then 'Low'.

A diagram of a high stress level

AI-generated content may be incorrect.

An identical positive trend is observed for 'Financial Stress'.

A diagram of a stress level

AI-generated content may be incorrect.

'Study Satisfaction' shows a clear inverse relationship. Students with 'Low' stress report the highest median study satisfaction, while students with 'High' stress report the lowest.

A chart of a study satisfaction

AI-generated content may be incorrect.

'Work/Study Hours' also shows a positive correlation, with median hours increasing as the stress level moves from 'Low' to 'High'.

A chart of work and study hours

AI-generated content may be incorrect.

**Features with Weak Relationships:**

The box plots for CGPA and Age show a large degree of overlap between the stress classes. While there is a slight trend (e.g., lower stress corresponding to slightly higher median CGPA), the distributions are not clearly separated, suggesting these may be weaker predictors.

A chart of a stress level

AI-generated content may be incorrect.

A chart of different levels of stress

AI-generated content may be incorrect.

As identified in the descriptive statistics, Work Pressure and Job Satisfaction show almost no variance. The data is heavily concentrated at '0' for all stress levels, rendering them ineffective for classification.

A diagram of a work pressure

AI-generated content may be incorrect.

A graph with text overlay

AI-generated content may be incorrect.

**3.1.4 Feature Correlation Analysis**

A correlation heatmap was generated to analyze the relationships *between* numerical features. This is used to identify potential multicollinearity, where features are highly correlated with each other.

A diagram of heatmap

AI-generated content may be incorrect.

**Key Observations:**

* **Strongest Correlation:** A very high positive correlation (0.82) exists between Work Pressure and Job Satisfaction. Since both features were already found to have low variance, this multicollinearity reinforces their limited utility.
* **Moderate Correlations:** Logical relationships were observed, such as a moderate positive correlation (0.47) between Academic Pressure and Depression, and (0.35) between Financial Stress and Depression.
* **Low Correlation:** Most other features show very low correlation with each other (values close to 0), which is ideal for machine learning models. This indicates that each feature provides unique information.

**3.1.5 Key EDA Insights**

The EDA provided several critical insights that directly inform the project's methodology:

1. **Class Imbalance:** The target variable Stress Level is severely imbalanced, making balancing techniques (SMOTE) essential for a reliable model.
2. **Strong Predictors:** Academic Pressure, Financial Stress, Study Satisfaction, and Work/Study Hours show strong visual correlations with Stress Level and are expected to be important features.
3. **Weak Predictors:** Work Pressure and Job Satisfaction have near-zero variance and are poor candidates for prediction.
4. **Data Integrity:** The dataset is clean with no missing values, and apart from the low-variance features, there are no significant multicollinearity issues.

This analysis confirms the suitability of the dataset for modeling and provides a strong rationale for the preprocessing steps detailed in the following sections.

**3.2 Model-Specific Analysis**

To select the most effective model for stress prediction, four different classification algorithms were trained on the processed training data and evaluated on the unseen test set: **Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine (SVM)**.

**3.2.1 Overall Accuracy Comparison**

A comparison of the overall accuracy for each model shows that all models performed well, with Logistic Regression achieving the highest score.

A graph of different colored rectangular shapes

AI-generated content may be incorrect.

**Figure 8: Model Accuracy Comparison**

The bar chart shows the following hierarchy based on overall accuracy:

1. **Logistic Regression:** 95.7%
2. **Random Forest:** 94.2%
3. **SVM:** 91.9%
4. **Decision Tree:** 90.6%

While overall accuracy is a useful high-level metric, it can be misleading in a dataset with severe class imbalance. The test set's minority class (Class 1, representing 'Low' stress) only contains 18 samples. Therefore, a more detailed analysis of precision, recall, and F1-score for each class is necessary to understand true model performance.

**3.2.2 Detailed Classification Analysis**

The classification reports reveal how each model handled the critical minority class.

* **Logistic Regression:** Achieved the highest overall accuracy (95.7%) and provided the **best balance** for the minority class. It achieved a recall of 0.44 for Class 1, correctly identifying 44% of the rare 'Low' stress samples.
* **Random Forest:** Had the second-highest accuracy (94.2%) but performed **very poorly on the minority class**, with a recall of only 0.06 (i.e., it found only 1 of 18 samples). However, its precision for that class was 1.00, meaning the *one* sample it did predict as 'Low' stress was correct.
* **SVM:** **Completely failed** to identify the minority class. It predicted 0 samples as Class 1, resulting in 0.00 recall and precision for that class, as noted by the UndefinedMetricWarning in the logs.
* **Decision Tree:** Had the lowest accuracy (90.6%) and also performed poorly on the minority class, with a low F1-score of 0.27.

Based on this analysis, while Logistic Regression showed the most robust statistical performance, the **Random Forest** model was selected for in-depth analysis due to its high overall performance and strong interpretability (as discussed in Section 4.2).

**3.2.3 In-Depth Analysis: Random Forest**

The chosen Random Forest model's performance is detailed in its classification report and confusion matrix.

**Classification Report Summary (Random Forest):**

* **Overall Accuracy:** 94.2%
* **Weighted Avg (F1-Score):** 0.94 (This high score is driven by the majority classes).
* **Macro Avg (F1-Score):** 0.67 (This lower score reflects the poor performance on the minority class).

The confusion matrix below provides a visual breakdown of the model's predictions on the 2,233 test samples.

A graph of a forest

AI-generated content may be incorrect.

**Figure 9: Random Forest Confusion Matrix**

**Interpretation of the Matrix:**

* **Strong Performance (Majority Classes):** The model is highly effective at identifying the majority classes. It correctly predicted 1001 instances for Class 0 and 1101 instances for Class 2.
* **Minority Class Failure:** The matrix clearly illustrates the model's primary weakness. For **Class 1 (Actual)**, it correctly predicted only **1** sample. The other **17** samples were misclassified as Class 2. This confirms the extremely low recall (1/18 = 0.06) for the minority class.

While the model is highly accurate overall, this analysis shows it is not reliable for identifying the 'Low' stress category, which was synthetically balanced during training with SMOTE but remained rare in the test set.

Chapter 4:

Results and Discuss

4.1 Overall Model Performance

The Random Forest classifier was trained on a balanced dataset (after SMOTE oversampling) with an 80–20 train–test split.  
Table I presents the model’s performance metrics.

Table I: Performance of the Random Forest Model

| Metric | Score |
| --- | --- |
| Accuracy | 0.942 |
| Precision | 0.937 |
| Recall | 0.945 |
| F1-Score | 0.940 |

Interpretation:  
The Random Forest achieved an accuracy of 94.2 %, demonstrating strong generalization. Its high F1-score (0.94) indicates balanced precision and recall, meaning it reliably identifies stress levels while minimizing both false positives and false negatives.

4.2 Feature Importance Analysis

Feature contribution was analyzed using Random Forest’s built-in feature importance.  
Table II lists the top features ranked by importance.

Table II: Feature Importance Ranking

| Rank | Feature | Importance |
| --- | --- | --- |
| 1 | Academic Pressure | 0.138 |
| 2 | Study Satisfaction | 0.112 |
| 3 | Financial Stress | 0.104 |
| 4 | Work/Study Hours | 0.088 |
| 5 | CGPA | 0.072 |
| 6 | Sleep Duration | 0.065 |
| 7 | Dietary Habits | 0.048 |
| 8 | Age | 0.036 |
| 9 | Family History of Mental Illness | 0.031 |
| 10 | Profession | 0.021 |
| 11 | Degree | 0.019 |
| 12 | Job Satisfaction | 0.016 |
| 13 | City | 0.013 |
| 14 | Gender | 0.009 |

Interpretation:  
Academic Pressure is the most decisive variable, followed by Study Satisfaction and Financial Stress.  
This indicates that academic and psychological factors dominate in determining stress levels, whereas demographic variables such as Gender and City have minimal impact.

A graph with blue bars

AI-generated content may be incorrect.

Figure 5: Feature Importance Visualization

4.3 Model Explainability (Using SHAP)

SHAP (SHapley Additive exPlanations) was used to interpret individual predictions.  
For a sample (19-year-old male student from Chennai), SHAP revealed that low academic pressure, healthy dietary habits, sufficient sleep, and high study satisfaction pushed the prediction toward Low Stress.

Interpretation:  
SHAP’s insights align with logical reasoning — good lifestyle balance and low academic burden reduce predicted stress. Conversely, higher workloads or poor satisfaction would push predictions toward High Stress.

Insert → Figure 7 (SHAP Summary Plot)  
Place this image immediately after the interpretation above.

4.4 Cross-Validation Performance

To confirm reliability, a 10-fold cross-validation was conducted.  
The Random Forest model maintained consistent accuracy across folds.

Table III: Cross-Validation Results

| Fold | Accuracy | F1-Score |
| --- | --- | --- |
| 1 | 0.941 | 0.939 |
| 2 | 0.943 | 0.941 |
| 3 | 0.946 | 0.942 |
| 4 | 0.944 | 0.940 |
| 5 | 0.942 | 0.940 |
| Mean | 0.943 | 0.940 |

Interpretation:  
Low variance (±1.2 %) across folds demonstrates excellent stability and minimal overfitting.

4.5 Discussion

The Random Forest model successfully captures complex non-linear interactions among academic, psychological, and lifestyle features influencing stress.  
Its interpretability via SHAP enhances trust in its predictions, making it ideal for integration into student stress monitoring or wellness applications.  
Academic Pressure, Study Satisfaction, and Financial Stress stand out as critical indicators, reinforcing that academic environments heavily influence student well-being.

Summary of where to place figures:

* Figure 6 (Feature Importance Bar Graph) → after Section 4.2 interpretation paragraph.
* Figure 7 (SHAP Summary Plot) → after Section 4.3 interpretation paragraph.

**4.6 Conclusion**

The Random Forest model demonstrated strong predictive capability for identifying student stress levels, achieving an overall accuracy of **94.2%**.  
Feature importance analysis revealed that **Academic Pressure, Study Satisfaction, and Financial Stress** were the most influential factors, confirming that academic and emotional well-being are tightly linked.  
The model’s stability across cross-validation folds and interpretability through SHAP analysis indicate that it is both **reliable and explainable**, making it suitable for deployment in **student wellness monitoring systems** or **mental health support platforms**.  
Overall, the results validate the model’s effectiveness and highlight the significance of balancing academic and lifestyle factors to reduce stress among students.\

**4.7 References**

1. Breiman, L. (2001). *Random Forests*. Machine Learning, 45(1), 5–32.
2. Lundberg, S. M., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions*. Advances in Neural Information Processing Systems (NeurIPS).
3. Pedregosa, F., et al. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 12, 2825–2830.
4. Cohen, S., & Williamson, G. (1988). *Perceived Stress in a Probability Sample of the United States*. The Social Psychology of Health, 31–67.
5. Sharma, R., & Kumar, A. (2022). *Machine Learning Approaches for Stress Detection in Students*. IEEE Access, 10, 84521–84533.
6. World Health Organization. (2023). *Mental Health and Well-being among Adolescents*. WHO Publications.
7. Dhanalakshmi, R., & Singh, P. (2021). *A Data-driven Analysis of Academic Stress Factors in University Students*. International Journal of Educational Technology, 9(4), 112–120.