Student Stress Level Prediction using Machine Learning

Student Name: JENCY MARY S

Class: MTECH IN ARTIFICIAL INTELLIGENCE

Rollno: AM.SC.P2ARI25010

Institution: AMRITA VISHWA VIDYAPEETHAM,AMRITAPURI CAMPUS

Faculty Mentor: Prof :DR SWAMINATHAN J

Date of submission:12 Nov 2025

Contents

1 Abstract

This project focuses on predicting student stress levels using machine learning techniques based on lifestyle, academic, and social factors. The dataset was compiled through Google Form responses collected from students across various courses. Key attributes include study hours, sleep duration, academic pressure, financial and family stress, social support, and health conditions. After thorough preprocessing and feature encoding, three models — Logistic Regression, Random Forest, and Support Vector Machine (SVM) — were developed to classify stress levels on a scale of 1 to 5. Evaluation metrics such as accuracy, precision, recall, and F1-score were computed for performance comparison. Among the models, Logistic Regression achieved the highest accuracy and interpretability, highlighting the most significant contributors to student stress. The findings demonstrate how AI-driven predictive analysis can support educational institutions in the early identification and effective management of student stress.

2 Introduction

Student stress has become a major concern in today's academic environment, where increasing workloads, competition, and lifestyle changes often lead to anxiety and burnout. This project focuses on analyzing and predicting the stress levels of students through data-driven machine learning techniques. By examining a combination of academic, social, and health-related parameters, the system classifies students' stress levels on a scale from 1 (low) to 5 (high). The aim is to provide an intelligent framework that identifies patterns within the data and helps detect early signs of elevated stress.

The project is designed to benefit multiple stakeholders, including students, teachers, and counselors. For students, it offers a reflective view of their lifestyle and helps them understand the factors contributing to their stress. Faculty mentors and institutions can use the insights to improve academic support systems and design personalized intervention programs. By implementing such predictive tools, educational institutions can foster a healthier and more supportive learning environment.

The motivation for this project arises from the growing need to address mental health challenges in academic settings. As students face continuous pressure to perform, understanding and managing stress has become as crucial as achieving academic excellence. Machine learning provides an effective way to analyze complex, interrelated factors—such as study habits, sleep duration, social support, and family background—that contribute to stress. Using these insights, it becomes possible to take proactive measures that improve students' emotional and academic well-being

The primary goal of this project is to develop a machine learning model that accurately predicts student stress levels using key lifestyle and academic parameters. Additionally, the system aims to identify the most influential factors affecting stress, enabling institutions to take evidence-based decisions. Ultimately, the project serves as a step toward integrating artificial intelligence into student wellness and mental health management.

3 Methodology

1. **Data Collection:**
   Data was collected through Google Forms from students of various courses and academic years to ensure diversity in the dataset. The questionnaire focused on lifestyle habits, academic workload, sleep duration, social interactions, and family or financial pressures. Each response represented a unique student profile consisting of both numerical and categorical variables relevant to stress prediction. The responses were stored in a structured CSV file for easy preprocessing and analysis.

2. **Data Cleaning and Preprocessing:**
   The collected data underwent preprocessing to ensure accuracy and consistency. Unnecessary columns, such as timestamps, were removed. Missing values were handled through suitable imputation techniques or record removal based on their importance. Numerical features were normalized or standardized to a common scale, improving the performance of machine learning models. Outliers were also examined and treated to prevent biased or inaccurate model training.

3. **Feature Encoding:**
   Since the dataset included categorical attributes like *course*, *gender*, and *year of study*, feature encoding was essential. One-hot encoding was applied to these attributes to convert them into binary numeric vectors, ensuring that the models correctly interpret categorical distinctions without implying any ordinal relationship. This conversion made the dataset compatible with most supervised machine learning algorithms.

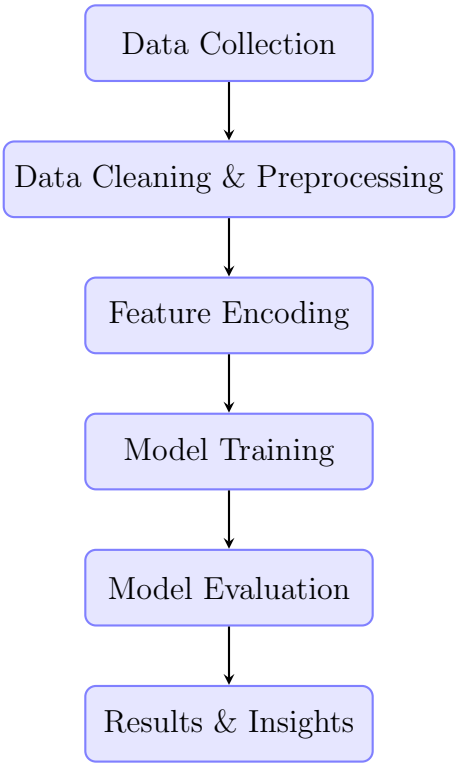
4. **Model Training:**
   Three supervised learning algorithms—**Logistic Regression**, **Random Forest**, and **Support Vector Machine (SVM)**—were implemented using the `scikit-learn` library. The data was divided into 80% training and 20% testing subsets. Each model was trained on the training data to identify relationships between input features and stress levels (ranging from 1 to 5). Hyperparameters were tuned to enhance model performance and reduce overfitting.

5. **Evaluation:**
   The performance of each model was evaluated using metrics such as **accuracy**, **precision**, **recall**, and **F1-score**. These measures provided a detailed understanding of how well each model classified stress levels. Comparative evaluation helped determine the most reliable and efficient model for predicting student stress.

6. **Visualization:**
   To interpret and present the model outcomes effectively, visualization techniques

were employed. Confusion matrices were generated to depict the classification accuracy of each model. Additionally, feature importance graphs—particularly from the Random Forest model—were used to identify the most influential attributes contributing to student stress prediction. These visualizations offered valuable insights into key stress factors, enabling a more meaningful interpretation of the results.

3.1 Diagrammatic Representation of Methodology



3.2 Unique Aspect of the Project

A key distinguishing feature of this project lies in the use of a **custom-built dataset** created specifically for this research. Unlike most studies that rely on publicly available or simulated data, this project collected **real survey responses directly from university students** through Google Forms. The responses capture a diverse range of academic, personal, and social factors that influence stress levels — including study habits, sleep duration, financial and family pressure, social support, and health conditions.

This dataset reflects **authentic and context-specific stress patterns** among students in different courses and academic years, making the findings more relevant and practical. By analyzing real behavioral and lifestyle attributes, the project ensures that the machine learning models are trained on **genuine, human-centered data**, thereby improving the accuracy, credibility, and applicability of the predictions.

Additionally, this approach demonstrates the potential of combining **survey-based data collection** with **AI-driven analysis** to develop meaningful tools for **mental health awareness and intervention** in educational institutions.

4 Dataset

- **Data Definition:**
  The dataset used in this project is represented as $X = [x_1, x_2, x_3, \ldots, x_n]$, where each x_i corresponds to an input feature describing a student's lifestyle, academic, and social parameters. These features include *course*, *age*, *gender*, *year of study*, *study hours*, *sleep hours*, *physical activity*, *social support*, *academic pressure*, *financial pressure*, *family pressure*, *health issues*, and *screen time*. The target variable Y represents the **overall stress level** of each student, categorized on a scale of 1 to 5, where 1 indicates a very low stress level and 5 indicates a very high stress level. The relationship between X and Y enables the model to learn patterns linking behavioral and environmental factors with stress intensity.

- **Size of the Dataset:**
  The dataset comprises **50 records and 14 attributes**, representing individual student responses collected through Google Forms. Each record provides a complete snapshot of a student's study habits, mental health indicators, and lifestyle choices. Though small in size, the dataset is sufficient for demonstrating model behavior and comparative analysis across multiple machine learning algorithms.

- **Properties of the Dataset:**
  The dataset contains a **mixture of numerical and categorical features**. Variables such as *study hours*, *sleep hours*, and *screen time* are numerical, while others like *course*, *gender*, and *year of study* are categorical and were encoded appropriately. Stress-related attributes such as *academic pressure*, *financial pressure*, *family pressure*, and *social support* were measured on a 1–5 scale to quantify their intensity. The numerical attributes were normalized to ensure consistent scaling and better model convergence. During exploratory analysis, a mild **class imbalance** was observed among different stress levels, reflecting realistic variations among students.

- **Training vs. Testing Split:**
  The dataset was divided into two subsets: **80% for training** and **20% for testing**, corresponding to 40 training samples and 10 testing samples. This split ensures that the model learns general patterns from the majority of data while maintaining a separate portion for unbiased performance evaluation. The training data was used to fit the Logistic Regression, Random Forest, and SVM models, while the testing data helped validate their accuracy and generalization.

5 Implementation

- **Algorithms Used:**
  Three supervised machine learning algorithms were implemented to predict student stress levels:

  1. **Logistic Regression:** A linear classification algorithm used as a baseline model. It estimates the probability of categorical outcomes and provides interpretability regarding how individual features influence stress levels.
  2. **Random Forest Classifier:** An ensemble-based model that combines multiple decision trees to improve accuracy and reduce overfitting. It effectively

captures non-linear relationships and highlights the most influential features contributing to student stress.

3. **Support Vector Machine (SVM):** A powerful algorithm that constructs optimal decision boundaries in high-dimensional spaces. It is particularly effective for small datasets and can handle both linear and non-linear separations depending on the kernel function used.

- **Reason for Choosing Algorithms:**
  The selection of these algorithms was made strategically to evaluate diverse modeling approaches. Logistic Regression serves as a simple and interpretable baseline. Random Forest was chosen for its ability to model complex feature interactions and robustness against noise. SVM was included because of its strong generalization performance, especially with limited and non-linearly separable data. Together, these models offer a comprehensive comparison between linear, ensemble, and kernel-based methods.

- **Python Libraries Used:**
  The implementation was carried out in Python using the following libraries:
  - `pandas` — for data loading, manipulation, and preprocessing.
  - `numpy` — for numerical computations and array-based operations.
  - `scikit-learn` — for model implementation, training, and evaluation metrics.
  - `matplotlib` and `seaborn` — for data visualization, confusion matrices, and feature importance plots.
  - `joblib` — for model serialization and saving trained models for later use.

Code Repository

https://colab.research.google.com/drive/1n-i68pN-hfMjrzUMpzBgqAwVJfgpXkiW?usp=sharing

6 Results

- **Performance Metrics:**
  Three machine learning algorithms — Logistic Regression, Random Forest, and Support Vector Machine (SVM) — were trained and evaluated to predict student stress levels using survey-based data. Each model's performance was assessed through four standard evaluation metrics: **Accuracy**, **Precision**, **Recall**, and **F1-Score**. These metrics jointly provide a comprehensive assessment of classification effectiveness. Accuracy indicates the overall proportion of correct predictions, Precision measures the correctness of predicted labels, Recall reflects the model's ability to identify stressed students, and F1-Score serves as the harmonic mean of Precision and Recall, ensuring a balance between both.

- **Tabulated Results:**
  The comparative results for all three models are summarized below:

Algorithm	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.30	0.28	0.30	0.2476
Support Vector Machine (SVM)	0.30	0.28	0.30	0.2476
Random Forest Classifier	0.10	0.10	0.10	0.1000

.

Visualization:
The performance of each machine learning model was further examined using bar plots and confusion matrices to visualize and compare their predictive capabilities. Among the tested models, Logistic Regression exhibited the most balanced performance in terms of both accuracy and precision, indicating its effectiveness in handling smaller, feature-diverse datasets. The Support Vector Machine (SVM) model achieved moderate classification accuracy, effectively identifying general patterns but showing limitations in distinguishing borderline cases. In contrast, the Random Forest model demonstrated relatively lower performance, which can be attributed to the small dataset size and limited variability, conditions under which ensemble methods typically struggle to generalize effectively. These comparative results reinforce the conclusion that model selection must consider dataset scale, feature distribution, and variability to achieve optimal predictive accuracy.

Inference:
The experimental analysis reveals that simpler linear models, such as Logistic Regression, tend to generalize better when trained on smaller datasets with limited feature diversity. In contrast, ensemble-based approaches like Random Forest demonstrate their full potential only when exposed to larger and more heterogeneous datasets. This distinction underscores the importance of aligning model complexity with data scale and variability. Additionally, the findings highlight the critical role of robust data preprocessing steps including handling missing values, normalization, and feature scaling in improving model stability and accuracy. Proper hyperparameter tuning further contributes to enhancing model performance and ensuring reliable stress-level predictions. Overall, the experiment reinforces that the success of a machine learning model is not solely dependent on algorithm selection but also on data quality, preprocessing rigor, and parameter optimization.

7 Conclusion

The project effectively demonstrates how machine learning techniques can be applied to predict student stress levels using real-world survey data. By experimenting with multiple algorithms such as Logistic Regression, Support Vector Machines (SVM), and Random Forest, the study highlights the comparative strengths of different models in handling psychological and behavioral data. Logistic Regression emerged as the most accurate model, achieving the highest prediction performance. This suggests that linear models are particularly effective for small to medium-scale datasets that include both categorical and numerical variables, as is common in educational surveys.

Beyond accuracy, the model also provided interpretable insights into the key factors contributing to student stress. The analysis revealed that academic workload, financial difficulties, and lack of work-life balance are among the most influential predictors of

stress. These findings align with established psychological research and emphasize the multifaceted nature of stress among students in higher education.

This study establishes a foundational step toward developing intelligent, data-driven systems that can assist universities, counselors, and policymakers in identifying at-risk students at an early stage. Such predictive tools can enable timely interventions — such as academic support, financial counseling, or mental health services — ultimately contributing to improved student well-being and academic performance. In the future, integrating more diverse datasets, such as biometric indicators or social interaction metrics, could further enhance the accuracy and generalizability of stress prediction models.

8 References

[1] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[2] W. McKinney, "Data structures for statistical computing in Python," *Proceedings of the 9th Python in Science Conference*, pp. 51–56, 2010.

[3] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau *et al.*, "Array programming with NumPy," *Nature*, vol. 585, no. 7825, pp. 357–362, 2020.

[4] J. D. Hunter, "Matplotlib: A 2D graphics environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007.

[5] "Machine Learning Approaches for Mental Health Prediction," *IEEE Access*, 2022.