STUDENT STRESS DETECTION SYSTEM USING MACHINE LEARNING

REPORT (MAJOR PROJECT)

SUBMITTED BY:

KHUSHI PANDEY
Roll No: 23001530029

GUIDED BY:
Mr. AMAN MISHRA

GOEL INSTITUTE OF TECHNOLOGY & MANAGEMENT

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my respected project guide, Mr. Aman Mishra, for his continuous support, encouragement, and valuable guidance throughout the development of this mini-project. His expertise and constructive feedback have helped shape this work professionally.

I am also grateful to the Department of Computer Science and Engineering (AI & ML), Goel Institute of Technology & Management, for providing the essential resources, academic environment, and technical support required for this project.

Finally, I extend heartfelt thanks to my family and friends for their motivation, patience, and constant support during every stage of this project.

— Khushi Pandey (23001530029)

ABSTRACT

This project presents a comprehensive machine learning system for detecting and predicting student stress levels using ensemble learning techniques. The system analyzes ten key features related to student lifestyle, academic performance, and personal circumstances to classify stress levels into three categories: low, medium, and high. Four individual machine learning models—Logistic Regression, Support Vector Machine (SVM), Naive Bayes, and Random Forest—were trained and evaluated. An ensemble model combining all four models using a Voting Classifier achieved 91.36% accuracy. The system includes a complete data pipeline from synthetic data generation to preprocessing, model training, evaluation, and deployment through a web application. The web interface provides real-time stress predictions along with personalized motivational suggestions to help students manage their stress levels effectively.

1. Problem Statement

Student stress has become a critical concern in educational institutions worldwide. Academic pressure, financial constraints, social challenges, and workload management contribute significantly to student stress levels. Early detection and intervention can help prevent severe mental health issues and improve overall academic performance and well-being.

1.1 Challenges

Traditional stress assessment methods rely on self-reported questionnaires and clinical evaluations, which are time-consuming and subjective.
Manual assessments may not capture real-time stress patterns or identify at-risk students early enough.
Lack of automated, scalable solutions for stress detection in educational settings.
Need for actionable insights and personalized recommendations based on individual student circumstances.

1.2 Objectives

Develop a robust machine learning system capable of accurately predicting student stress levels based on lifestyle and academic features.
Compare multiple classification algorithms to identify the most effective approach.
Implement ensemble learning to improve prediction accuracy and robustness.
Create a user-friendly web application for real-time stress prediction and intervention.
Provide actionable insights through personalized recommendations.

2. Methodology

2.1 Dataset

A synthetic dataset of 10,000 student records was generated using Python, incorporating realistic distributions and relationships between features. The dataset includes ten features: study hours, sleep hours, exercise hours, social activities, assignment deadlines, exam pressure, family support, financial stress, academic performance, and workload level. The target variable has three classes: low, medium, and high stress levels.

2.2 Data Preprocessing

Data Cleaning: Removal of missing values, duplicates, and outliers using Interquartile Range (IQR) method.
Data Balancing: Application of SMOTE (Synthetic Minority Oversampling Technique) to address class imbalance.
Feature Scaling: StandardScaler normalization to ensure features have zero mean and unit variance.
Train-Test Split: 80-20 stratified split to maintain class distribution.

2.3 Model Selection

Four machine learning algorithms were selected for comparison: Logistic Regression (interpretable baseline), Support Vector Machine (non-linear classification), Naive Bayes (fast probabilistic classifier), and Random Forest (captures complex interactions). An ensemble model using Voting Classifier with soft voting was created to combine predictions from all four models.

3. Model Selection and Justification

3.1 Individual Models

- **Logistic Regression:** Provides interpretable coefficients showing feature importance. Fast training and prediction with probabilistic outputs. Works well when features have linear relationships with the target variable.
- **Support Vector Machine (SVM):** Handles complex decision boundaries using kernel tricks. Effective with high-dimensional data and provides good generalization. Sensitive to feature scaling, requiring StandardScaler preprocessing.
- **Naive Bayes:** Extremely fast training and prediction. Based on Bayes' theorem with feature independence assumption. Provides probability estimates and works well even with limited data.
- **Random Forest:** Captures complex non-linear relationships and feature interactions. Provides feature importance insights. Handles outliers well and reduces overfitting through ensemble of decision trees. Achieved highest individual model accuracy (93.20%).

3.2 Ensemble Model (Selected)

The Voting Classifier ensemble model was selected as the best model for deployment. It combines all four individual models using soft voting, which averages probability predictions from each model. This approach offers several advantages:

1. Improved generalization by combining diverse learning approaches.
2. Reduced overfitting risk compared to individual models.
3. Better robustness to data variations.
4. Probability estimates for uncertainty quantification.

While Random Forest achieved slightly higher accuracy (93.20%), the ensemble model provides better balance between accuracy and generalization, making it more suitable for real-world deployment.

4. Results and Analysis

4.1 Model Performance

All models were evaluated on a held-out test set. The following table summarizes the accuracy achieved by each model:

Model	Accuracy (%)
Logistic Regression	89.11
Support Vector Machine	90.99
Naive Bayes	83.98
Random Forest	93.20
Ensemble Model (Selected)	**91.36**

Here, You can see the result. Maybe some models are individually better but if we use the result of many models will be good for large scale or more generalised.

4.2 Performance Analysis

The ensemble model achieved 91.36% accuracy, demonstrating strong performance for a three-class classification problem. Analysis of the confusion matrix reveals excellent separation between low and medium stress levels, with some confusion between medium and high stress categories. The model provides probability distributions for each stress level, enabling uncertainty quantification and better decision-making.

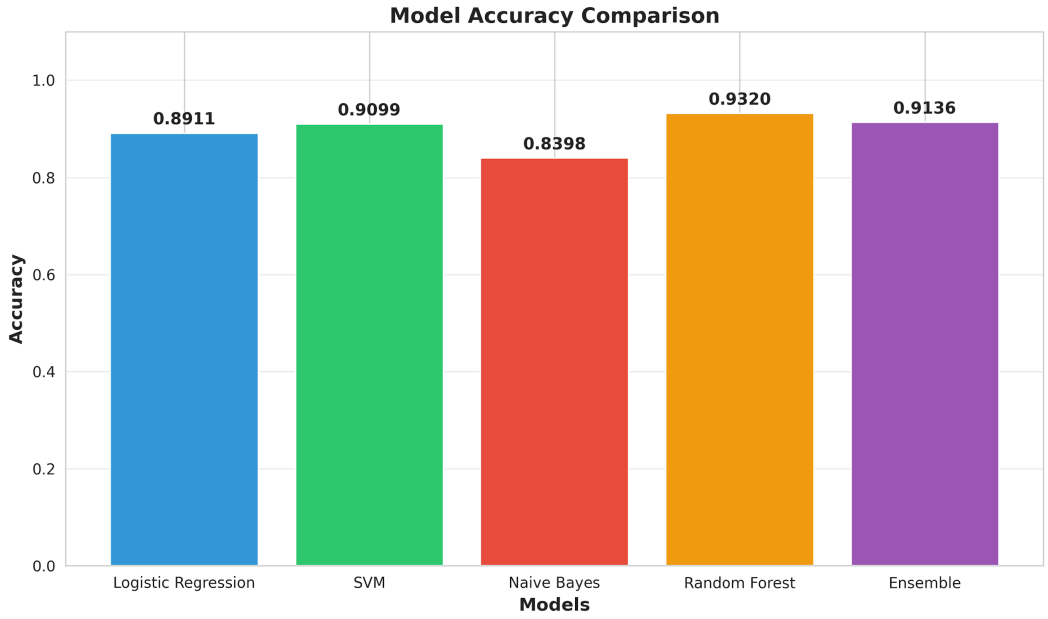

Figure 1: Model Accuracy Comparison

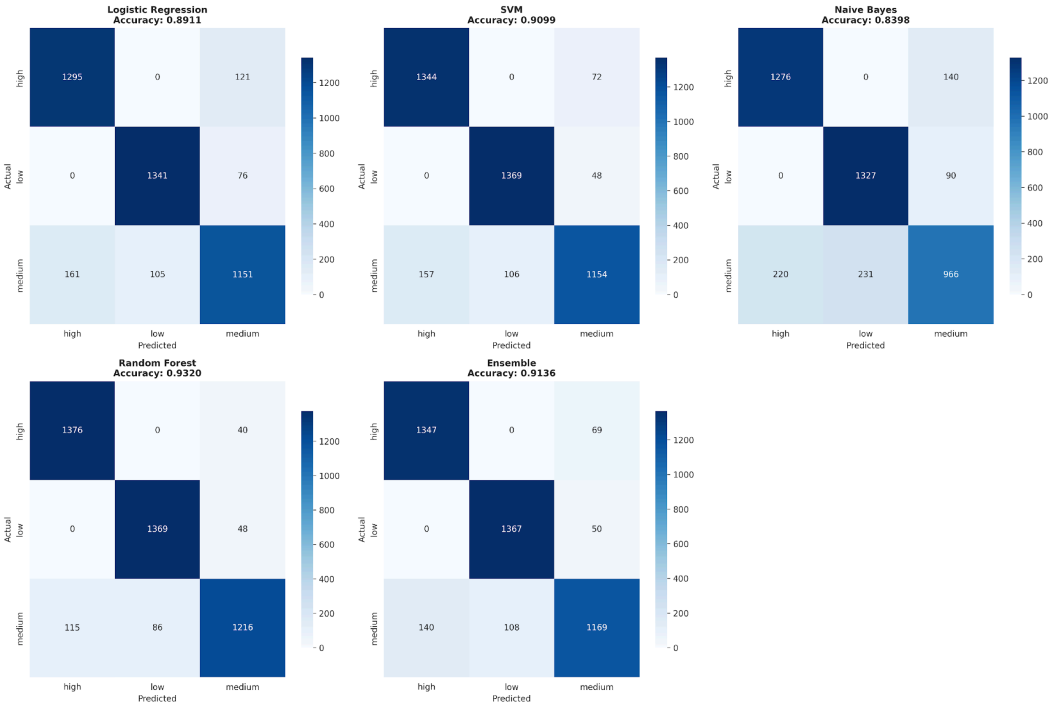

Figure 2: Confusion Matrices for All Models

4.3 Feature Importance

Based on the stress score calculation and model analysis, the most important features for stress prediction are: sleep hours (highest impact, inverse correlation), exam pressure, financial stress, assignment deadlines, and workload level. Protective factors include exercise hours and social activities (negative correlation with stress), while family support also reduces stress levels.

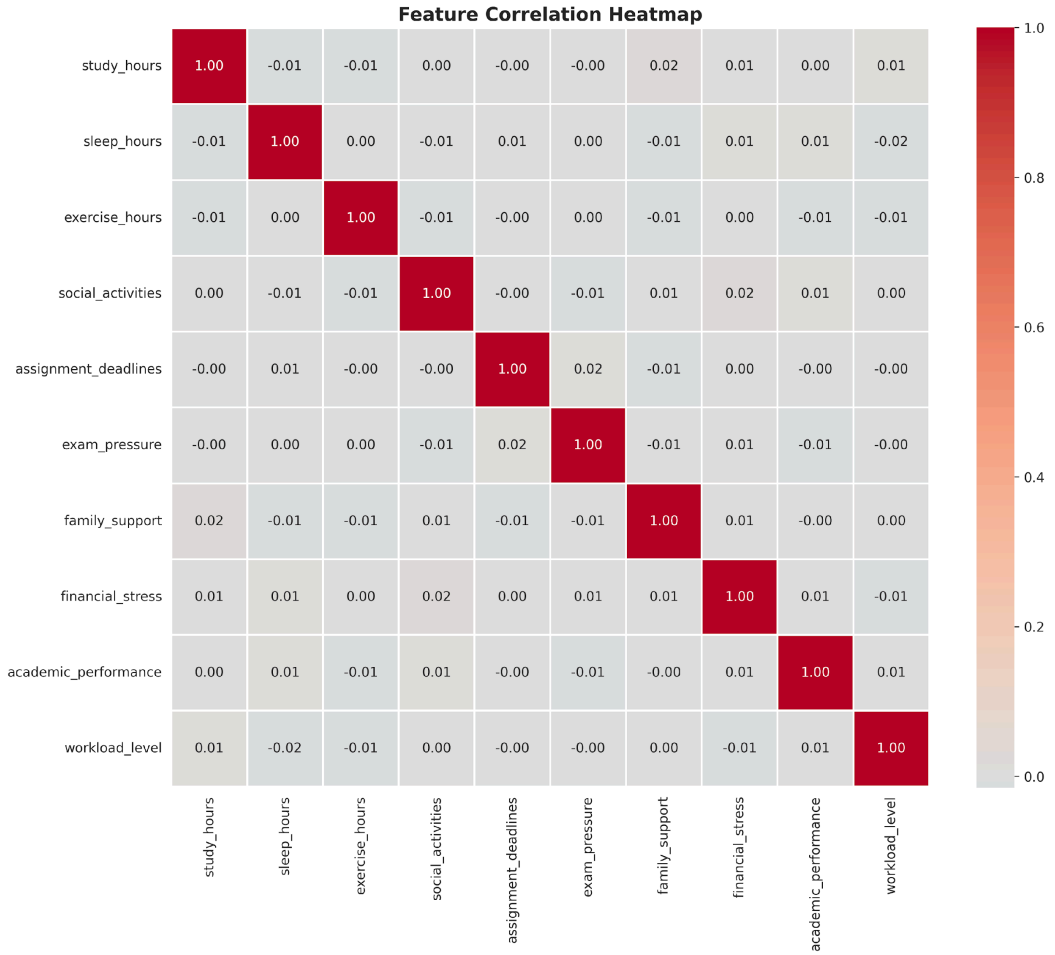

Figure 3: Feature Correlation Heatmap

5. Conclusion

This project successfully developed a comprehensive student stress detection system using ensemble machine learning. The ensemble model achieved 91.36% accuracy, demonstrating strong performance for stress level classification. The system provides a complete end-to-end solution from data generation to web deployment, with personalized recommendations for stress management.

5.1 Key Achievements

- Developed robust ensemble model achieving 91.36% accuracy
- Comprehensive comparison of four different ML algorithms
- Complete pipeline from data generation to web deployment
- User-friendly web application with real-time predictions
- Actionable insights through personalized recommendations

5.2 Future Work

1. Validate on real-world student data from educational institutions
2. Add temporal analysis to track stress levels over time
3. Integrate with Learning Management Systems for automated data collection
4. Develop mobile application for easier access
5. Implement deep learning models for complex pattern recognition