Module Code & Module Title

CU6051NI Artificial Intelligence

Individual Coursework

Submission: Milestone 1

Academic Semester: Autumn Semester 2025

Credit: 15 credit semester long module

Student Name: Sanskriti Agrahari

London Met ID: 23048503

College ID: NP01AI4A230002

Assignment Due Date: 21/01/2026

Assignment Submission Date: 21/01/2026

Submitted To: Er. Roshan Shrestha

GitHub Link	https://github.com/03sans/Student_Stress_Monitoring

I confirm that I understand my coursework needs to be submitted online via MST Classroom under the relevant module page before the deadline for my assignment to be accepted and marked. I am fully aware that late submissions will be treated as non-submission and a mark of zero will be awarded

Table of Contents

Table of Figures

Table of Tables

1. Introduction

1.1 Overview of the AI Topic

This coursework focuses on Supervised Machine Learning, especially on classification techniques on structured survey data. Supervised learning is one of the widely used approaches in machine learning where a model is trained with a labelled dataset. A known output label pairs each instance of an input. During the training, the model comprehensively learns the patterns and relationships between the input features and target variable so that it generalizes to make right predictions on unseen data. (IBM, n.d.).

Classification represents a basic category of supervised learning that assigns observations into predefined classes using learned patterns present in data (IBM, n.d.). In this coursework, classification is used to categorize students into different levels of stress, such as low, moderate, or high, based on various academic, behavioural, health-related, and lifestyle factors captured through a survey. Apart from the prediction of continuous values, as in regression tasks, classification problems predict discrete outcomes, making it particularly suitable for decision-support systems in educational settings.

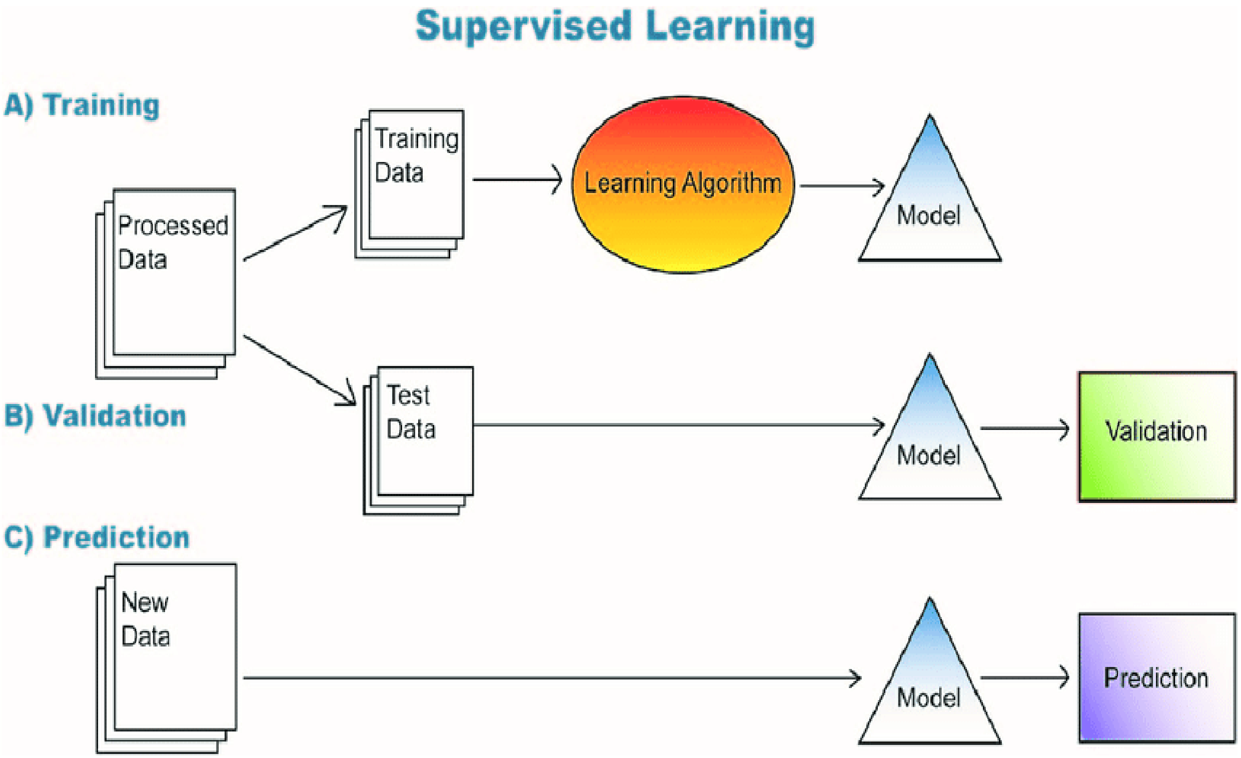


Figure 1 Graphic representation of supervised machine learning (Kanevsky, n.d.)

Supervised classification algorithms have an increasingly very core application within predictive analytics applications involving areas where early warning regarding risk is an imperative. In the education sector, applications involving studying students' engagement, predicting performance, analysing possible dropouts, as well as analysing possible concerns related to psychological or behavioural issues, involve supervised classification algorithms. These applications enable an institution to move from reactive approaches where issues are tackled only when they go bad to proactive approaches involving early intervention for those issues via learned patterns.

In this coursework, data cleaning, pre-processing, feature selection and encoding, model training, and performance evaluation using appropriate metrics are some of the major machine learning processes that form a conceptual foundation for the proposed system. These processes facilitate a structured design for a reliable student's stress classification framework.

1.2 Problem Domain: Student Stress Analysis

Stress among students has emerged as a serious concern in modern educational settings, with a particular emphasis on higher educational institutes. Higher work pressures in academics, ensuring good performance in examinations, poor time management skills, lack of proper sleep patterns in students, concern for health and related improper practices regarding a healthy lifestyle in students, are some of the prime factors that result in students experiencing stress. Students may face very serious effects in case they continue to suffer from high levels of stress in life.

Notwithstanding its pervasiveness, stress among university students can often pass undetected or be treated only when it becomes more serious. Conventional ways to determine when there is too much student stress may come from personal surveys, therapy sessions, or general observation conducted by instructors and university administrators. Even if these methods are very good, they can be somewhat unreliable, nonspecific, and subject to the willingness of the university student, as well as university personnel, to cooperate.

Additionally, the experience of student stress is a multi-dimensional phenomenon, which is the result of a complicated process of academic, social, behavioural, or health-related facets. For example, if the measure of student stress is based on some indications or manual assessment, it will be highly difficult to ascertain these phenomena accurately.

Therefore, a rising need is emerging for a data-intensive and automated system that has the ability to process various risk factors together to detect early-risk individuals for high-stress levels among university students.

1.2.1 Problem Statement

The coursework at hand, examines the problem which exists with regards to there being no legitimate means of analysing which of the students are experiencing a substantial degree of stress on a consistent basis to date. In fact, what currently occurs seems to be unorganized in a manner which results in these tactics proving ineffective for purposes of early intervention.

In an effort to move beyond these perceived inadequacies, the proposed project will utilize a method of classification by supervised machine learning that uses structured data gathered among students to forecast levels of stress. Rather than looking at each factor separately with distinct systems for analysis, combining elements of academics, behaviours, health, and lifestyle into one analytical system will assist in allowing for a more accurate assessment of stress levels by not being so narrow in focus.

1.3 Aim of the Proposed Study

The aim of this study is to propose a supervised machine learning-based classification system that uses student survey data to predict stress levels. By analysing academic, behavioural, and health-related attributes, the system aims to achieve:

- The identification of students at high risk for high levels of stress at an early stage.
- To provide support for educators and administrations with insights that are data informed.
- To contribute to improving the well-being of students, ultimately helping them so they can excel academically.

The coursework focuses on conceptual system design, algorithm selection, and logical workflow representation rather than full system implementation.

1.4 Dataset Overview

This study uses a student stress survey dataset obtained from the Hugging Face dataset repository *(0xmarvel/student-stress-survey).* The dataset contains responses collected through a structured questionnaire designed to capture multiple factors related to student stress.

The dataset includes 1,000 student responses and approximately 39 attributes covering academic resources, study environment, learning habits, health conditions, sleep patterns, lifestyle choices, and perceived stress indicators. One stress-related survey response is treated as the target variable, while the remaining attributes serve as independent features.

The column StudentStressSurvey_Id is used solely as a unique identifier and is excluded from model training, as it does not carry meaningful predictive information. The dataset is suitable for supervised machine learning because the independent variables represent real-world factors that logically influence student stress levels.

Attribute	Description
Dataset Source	Hugging Face (0xmarvel/student-stress-survey)
Number of Records	1,000 student responses
Number of Attributes	39 features
Data Type	Structured survey data
Feature Categories	Academic, behavioural, health, sleep, lifestyle, stress indicators

Target Variable	Stress-related survey response
Identifier Column	StudentStressSurvey_Id (excluded from training)
Suitability for ML	High features have logical influence on stress levels

Table 1 Dataset Overview

This dataset provides a reliable foundation for designing and analysing a supervised classification solution for student stress prediction.

2. Background

2.1 Student stress as an educational and wellbeing problem

Student stress has come into prominence in recent years as a key concern within the context of higher education because of the profound effects it exerts on students' academic outcomes. Recent studies with broad evidence have confirmed that the prevalence of students suffering from stress-related mental health problems has been found to be quite high across the world, making it difficult to consider it as a problem limited to regional or institutional settings (Paiva, et al., 2025). Studies rooted within the context of public health have also provided evidence pertaining to the profound

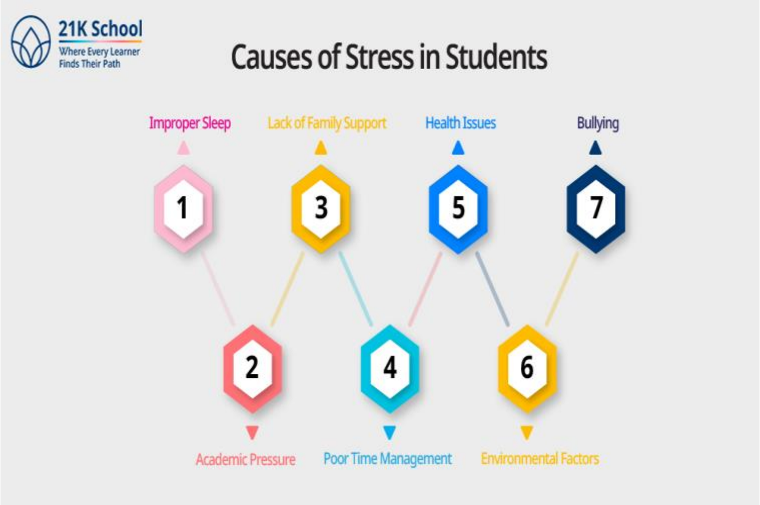


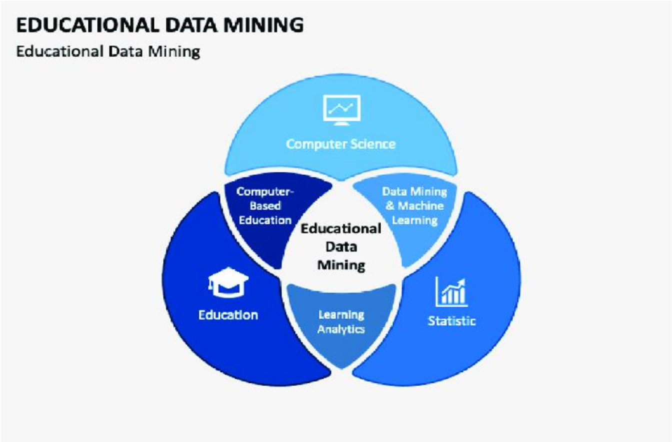
Figure 2 Common factors contributing to stress in students. (21kschools, 2025)

concern of students suffering from moderate to severe mental distress influenced by academic and behavioural factors (Malebari, et al., 2024).

From an issue domain perspective, stress among students is inherently multi-dimensional. Stress among students can result from several factors, with these factors inter-playing with each other in complex ways. These factors include, but are not limited to, study schedules, exam schedules, sleep patterns, physiological conditions, habits, and social conditions. Given these complexities, stress measurement using observation or factor-testing approaches may not be very accurate. This is because more precise stress measurement requires an integrated model.

2.2 Research trend: educational data mining and learning analytics

In the previous decade, the amount of data produced by educational institutions regarding students has been substantial. This data has been mainly of the type that is collected through surveys, learning/academic records, engagement, and indicators of students' wellbeing. A new field of research that focuses on the study of such data to improve understanding and decision-making regarding learning is Educational Data Mining and Learning Analytics (Papadogiannis, et al., 2024). In this context, supervised machine learning methods are often used for prediction tasks such as the identification of at-risk students, the prediction of academic results, or the implementation of subsequent intervention measures. In a survey of the current state of research in the field of predictions in the context of educational institutions, it was found that machine learning algorithms are often able to anticipate potential warning indications before conventional measurement processes are carried out, allowing for earlier academic or pastoral intervention (Li, et al., 2024). This encourages the application of



comparable prediction techniques to stress-related consequences, where early warning is crucial.

2.3 Machine learning for stress detection and prediction

Current research trends involving stress prediction via machine learning can essentially be categorized into two distinct domains. First, there is stress detection through sensors, where physiological data is used from wearable technologies. Current systematic reviews have authenticated its efficacy for stress pattern recognition, despite certain limitations identified for generalisability or validation (Pataca, et al., 2025).

In the second category, the predictions are made based on survey data and behaviour. Although the results from the sensor-based prediction systems are highly accurate, in the context of academic institutions, the cost of implementation, privacy issues, and the lack of accessibility are some of the limitations. In the case of survey data, the data sets like the student stress survey provide a better alternative for academic institutions, as this data set entails multiple aspects of stress, which are related to the subjects of bedtime routine, workload, health status, habits, and academic pressure.

2.4 Existing work in the student-stress prediction domain

Recent studies increasingly propose supervised learning frameworks that classify student stress using multi-factor inputs and categorical outputs such as low, moderate, or high stress levels. For example, context-aware machine learning frameworks based on survey data demonstrate that stress classification performance improves when diverse contextual features are combined and multiple algorithms are evaluated rather than relying on a single model (Ovi, et al., 2025).

In a similar manner, Applied Research often involves comparing the capabilities of models like Logistic Regression, Decision Trees, Random Forest Models, Support Vector

Machines, and Neural Networks in the creation of stress predictor tools for students (Yeler & Sürücü, 2025). These models tend to be preferred for the said researchers for the reason that they are capable of processing complex interactions between features while still being able to interpret results. The common process from the literature entails the execution of data cleaning procedures, encoding, feature scaling when necessary, splitting the data into training and test sets, model training, and the utilization of metrics like accuracy, precision, recall, and confusion matrices.

2.5 Explainability, ethics, and practical deployment considerations

In the applications concerning the well-being of students, transparency is a significant requirement. This is due to the fact that predictions from machine learning models can impact academic support intervention, counselling, and institutional decision-making; thus, stakeholders involved, such as teachers, counsellors, and students, need to interpret the predictions from machine learning models. Studies involving explainable artificial intelligence (XAI) in mental healthcare applications make it clear that interpretability strategies, such as feature importance techniques and model explainability techniques, can improve transparency (Tariq, et al., 2025).

Attention needs to be given to the ethical issues present, given the sensitive nature of student stress data. Systems need to protect privacy, ensure consent, and present predictive framing in a way that student data is used for their benefit rather than becoming a marker for stigma. Again, these points reiterate the need for conceptual solutions based upon a more explanation-oriented, ethical approach, rather than making actual predictive decisions through comprehensive, unaltered automated processes.

2.6 Review of Related Research Work

There has been a growing body of research that have investigated the implementation of machine learning techniques in order to analyse stress in students, their mental wellbeing, and even academic risk using survey-based and educational datasets. These papers offer crucial information about suitable datasets, algorithmic approaches, and expected outcomes for stress prediction and early risk identification in educational environments.

Research 1: (Paiva, et al., 2025)

- **Dataset:** Meta-analysis of global university student mental health and wellbeing survey studies
- **Algorithm:** A large-scale statistical analysis combined with machine learning-based synthesis methods

An extensive meta-analysis of the mental health and wellbeing of international university students was conducted in this research. This analysis focused on the aspects of stress and psychological issues in different geographical and educational environments through statistical analysis along with the application of machine learning-based synthesis approaches. The analysis indicated the presence of ongoing issues of student stress, as the problem is not limited to any environment. The researchers of this analysis strongly suggested the importance of prediction models in educational environments in the context of the importance of early detection systems to identify stress before it affects educational performance.

Research 2: (Malebari, et al., 2024)

- **Dataset:** Public health datasets focusing on student psychological distress, lifestyle factors, and academic pressure
- **Algorithm:** Predictive statistical modelling and risk-factor analysis

This research conducted research on public health data includes factors such as psychological disturbances among students, personal lifestyle factors, and academic pressure through predictive statistical modelling and risk factors. The research proved robust links between academic load, disrupted sleep patterns, and moderate to high levels of psychological disturbances among students. The research proved stress among students is determined by a combination of behavioural and personal factors and not a single one. The research therefore proves the need to use data comprising multiple factors and surveys when designing stress modelling systems.

Research 3: (Li, et al., 2024)

- **Dataset:** At-risk student datasets collected from educational institutions, including engagement and wellbeing indicators
- **Algorithms:** Supervised machine learning classifiers

In this study, they analysed student data from at-risk students gathered from institutions of learning, including factors like level of engagement and wellbeing metrics. The study used supervised machine learning algorithms to develop early warning signs of student risk. The findings of the study revealed that machine learning algorithms have been more effective than conventional risk assessment tools for spotting at-risk students for prospects of poor performance or mental wellness problems. This study illustrated how predictions can facilitate early interventions for academics and wellness support.

Research 4: (Ovi, et al., 2025)

- **Dataset:** Survey-based student stress datasets incorporating academic, behavioural, and contextual variables
- **Algorithms:** Logistic Regression, Support Vector Machine (SVM), and Random Forest

This study performed a comparative analysis of multiple supervised learning algorithms, including Logistic Regression, Support Vector Machines, and Random Forest, using survey-based student stress datasets. The dataset incorporated academic, behavioural, and contextual variables. The study found that combining diverse contextual features significantly improved stress classification accuracy. The authors highlighted the importance of feature diversity and algorithm comparison when designing machine learning-based stress prediction systems, particularly in survey-driven educational data.

Research 5: (Yeler & Sürücü, 2025)

- **Dataset:** Student performance and stress-related datasets
- **Algorithms:** Decision Trees, Random Forest, and Neural Networks

The work in this research examined student performance and stress-related datasets employing Decision Trees, Random Forest algorithms, and Neural Networks. The work showed that Random Forest performed well in terms of prediction but provided a high degree of explainability as well. The combination of both was considered to be of great use in making decisions in educational institutions, where explainability of algorithms was a key requirement.

All things considered, these studies demonstrate the efficacy of supervised machine learning techniques for predicting student stress, especially when employing survey-based datasets with multi-dimensional features. The design decisions in this coursework, such as the choice of dataset, algorithm, and evaluation method, are directly influenced by the results of previous research.

2.7 Comparative analysis of the research works

Study	Dataset Type	Algorithms / Methods Used	Key Focus	Key Findings	Relevance to This Coursework
Research 1	Global university mental health and wellbeing	Statistical analysis and ML-based synthesis	Understanding prevalence of student stress at a global level	Student stress is widespread and persistent	Justifies the importance of building early stress detection

	survey studies (meta-analysis)			across regions; early detection is critical	systems using data-driven approaches
Research 2	Public health datasets on student distress, lifestyle, and academic pressure	Predictive statistical modelling, risk-factor analysis	Identifying contributors to psychological distress	Stress is influenced by multiple academic, behavioural, and lifestyle factors	Supports the use of multi-dimensional survey datasets rather than single-factor analysis
Research 3	At-risk student datasets with engagement and wellbeing indicators	Supervised machine learning classifiers	Early identification of at-risk students	ML models outperform traditional screening methods	Reinforces the choice of supervised learning for early stress and risk prediction
Research 4	Survey-based student stress datasets with contextual variables	Logistic Regression, SVM, Random Forest	Algorithm comparison for stress classification	Feature diversity improves classification accuracy	Motivates algorithm comparison and feature engineering used in this coursework

Research 5	Student performance and stress-related datasets	Decision Trees, Random Forest, Neural Networks	Balancing performance and interpretability	Random Forest provides strong accuracy with interpretability	Supports inclusion of tree-based models for explainable educational decision support
This Coursework	Survey-based student stress dataset	Logistic Regression, Decision Tree, Random Forest	Multi-class stress prediction and model comparison	Traditional ML models provide interpretable and effective stress classification	Builds directly on prior research while focusing on practical implementation and evaluation

Table 2 Comparative Analysis of Research Done

2.8 Summary of research gap and justification for this coursework

Overall, research shows strong potential for machine learning to classify or predict stress, and educational analytics research supports using supervised models for early risk detection (Papadogiannis, et al., 2024) (Li, et al., 2024). However, gaps remain in building practical, survey-based stress prediction pipelines that handle mixed categorical/numerical features robustly, clearly defining labels from survey instruments in a consistent way, and presenting solutions with transparent logic suitable for real academic settings (Tariq, et al., 2025).

Building upon the conceptual analysis conducted in Coursework 1, this coursework addresses these gaps by proposing a supervised classification approach using a structured student stress survey dataset. The study focuses on developing a clear conceptual pipeline that includes justified algorithm selection, systematic preprocessing, model training and evaluation, and diagrammatic representation. This approach aligns with realistic educational needs while emphasising ethical considerations, explainability, and decision-support relevance.

3. Solution

3.1 Overview of the Proposed Solution

The implemented solution is a supervised machine learning classification system that predicts a student's stress level using structured survey responses from the Student Stress Survey dataset (Hugging Face: 0xmarvel/student-stress-survey). The dataset contains multiple attributes representing students' academic situation, habits, wellbeing, sleep and lifestyle conditions, and other stress-related indicators. The system treats the survey question "How Would You Rate Your Stress Level During This Academic Term" as the target label, while all remaining survey attributes are used as input features.

A complete end-to-end machine learning pipeline was developed and tested. The workflow begins by loading the dataset and performing basic label cleaning (e.g., trimming inconsistent spacing/casing and removing empty/blank class values). After cleaning, the target stress label is converted into numeric form using Label Encoding, enabling it to be used by machine learning models. Since several survey attributes are categorical, the feature set is transformed using One-Hot Encoding (pd.get_dummies) to convert categorical responses into binary indicator variables, producing a fully numerical dataset suitable for model training.

The data is then split into training and testing sets using an 80/20 train–test split, with stratification applied to preserve the original stress-class distribution in both subsets. Because the dataset contains class imbalance (some stress categories appear more frequently than others), the implemented models apply class weighting (class_weight="balanced") so that minority classes receive higher importance during training.

Three supervised classification models were implemented and compared:

- Logistic Regression (baseline model)
- Decision Tree Classifier (non-linear, rule-based model)
- Random Forest Classifier (ensemble of trees to reduce overfitting and improve generalisation)

In addition to the baseline Random Forest, a tuned Random Forest configuration was also trained using improved hyperparameter settings to improve predictive performance and reduce misclassification.

Model performance was evaluated on the test set using standard classification measures, including:

- Classification Report (precision, recall, F1-score per class)
- Confusion Matrix (to analyse misclassification patterns across Low/Moderate/High stress categories)
- Overall Accuracy and F1-scores (macro and weighted for balanced comparison)

Finally, the implemented solution includes testing on unseen/new survey inputs by encoding new input records, aligning their encoded feature columns to match the trained model's feature space, generating predictions, and converting predicted numeric labels back to human-readable stress categories using the inverse label encoder transformation.

Overall, the implemented solution demonstrates a full supervised machine learning pipeline from raw survey data through preprocessing, model training, evaluation, and prediction capable of classifying student stress into interpretable categories that can support early identification and potential interventions in educational settings.

3.2 System Workflow and Implemented Approach

The implemented solution follows a structured supervised machine learning workflow, where each stage transforms the raw student survey data into reliable and interpretable stress-level predictions. The workflow ensures data quality, fair learning despite class imbalance, and meaningful evaluation of multiple models.

Stage 1: Dataset Loading and Initial Analysis

Firstly, the data workflow introduces the student stress survey data to the Python platform with the use of the Pandas library. At the outset, the data survey is undertaken to determine the data structure, the types, and the variations in the responses concerning the level of stress. It is crucial in identifying challenges such as mismatched labels, unused identifier variables, and class imbalance.

In this phase, the unique identifier column of the survey is removed because it doesn’t add much in terms of predictive learning; rather, it can cause some noise in the model.

Stage 2: Target Label Cleaning and Encoding

The variable with the target, which is the stress level of the students, is proceeded to clean. White space from the leading or trailing, misspellings in the text formatting, and missing values in the class are removed. This ensures that the values of the stress level (Low, Moderate, High) are unique.

Following this, the categorical values of stress are converted to numerical values through Label Encoding. This makes it easier for supervised learning algorithms to work with the target value and also ensures that a numerical value is mapped to a specific value of stress.

Stage 3: Feature Encoding and Dataset Preparation

Most attributes in the survey are categorized. For handling these attributes using machine learning, One-Hot Encoding is used via pd.get_dummies(). Using this, the attributes are coded in such a way that their responses are converted into indicators, without any artificial ordering among the variables.

The encoded data now has only numeric attributes and can be processed by all classification algorithms selected for classification. This process causes an abrupt rise in dimensionality but keeps the data significance and interpretability of responses in the survey form intact.

Stage 4: Train-Test Split with Class Stratification

The prepared dataset is then divided into a training set and a testing set using an 80/20 ratio. Stratified splitting techniques are used to ensure equal proportion of all categories of stresses in the divided dataset.

This is more necessary since there is class imbalance in the data. This means certain stress levels are more common in the dataset than others. The class distribution needs to be maintained since it will help in an unbiased assessment of the model.

Stage 5: Model Training with Class Balancing

Three supervised classification models are trained on the prepared training dataset:

- Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier

For balancing the distribution in stress classes, class weighting is used while training, where more emphasis is given to the classes using the parameter: “class_weight = 'balanced'”. It gives more emphasis to the small classes, ensuring the network is not biased toward the dominant class because of their higher counts in testing sets due to their higher occurrence in the training data sets.

In each model, the association between the encoded survey attributes and the stress label is learned. This helps the model develop decision boundaries for classifying stress into the categories of Low, Moderate, and High.

Stage 6: Model Evaluation and Performance Analysis

After training, each model is evaluated using the test dataset. Performance is measured using:

- **Accuracy** to assess overall correctness
- **Precision and recall** evaluating class-specific performance
- **F1-score** to balance precision and recall
- **Confusion Matrix** to analyse misclassification patterns between stress categories

This evaluation allows direct comparison between baseline and more complex models, highlighting trade-offs between interpretability and predictive performance.

Stage 7: Hyperparameter Tuning and Model Refinement

In addition to improving the model’s performance, a tuned model of Random Forest is developed by using improved hyperparameters, for instance by increasing the number of trees in addition to limiting tree depth and sample sizes.

The performance of the tuned model is assessed on the same criteria so that comparisons can be drawn.

Stage 8: Prediction on Unseen Data

In the final stage, the trained model is used to predict stress levels for new, unseen student survey inputs. New input data is encoded using the same preprocessing and feature encoding steps applied during training. Feature alignment is performed to ensure consistency between training and prediction feature spaces.

Predicted numeric labels are then converted back into human-readable stress categories using the inverse label encoding process. This stage demonstrates the system's practical applicability in real educational scenarios.

3.3 AI Algorithms Used

The implemented solution applies multiple supervised classification algorithms to predict student stress levels from structured survey data. All selected algorithms are well-suited for categorical and numerical survey features and are commonly used in educational and wellbeing analytics. Implementing multiple models enables a comparative evaluation, allowing strengths and limitations of each approach to be analysed before selecting the most suitable model.

To address class imbalance present in the dataset, class weighting (class_weight="balanced") is applied during training wherever supported. This ensures that minority stress categories receive adequate importance during the learning process.

While the initial experiments were conducted using a five-class stress formulation, further analysis later in this study motivated a reformulation into three ordinal stress categories to improve stability and real-world interpretability.

3.3.1 Logistic Regression

Logistic Regression is implemented as a baseline classification model. It estimates the probability of each stress class using a linear combination of the input features, followed by a SoftMax function for multi-class classification. Despite its simplicity, Logistic Regression provides strong interpretability, as feature coefficients indicate the direction and relative influence of individual survey attributes on predicted stress levels.

In this implementation, Logistic Regression serves as a benchmark to evaluate whether more complex models provide meaningful performance improvements. Class weighting is applied to reduce bias toward majority stress categories. Even though the Logistic Regression algorithm runs very efficiently and is easy to interpret, its linear decision boundaries make it less capable of capturing the non-linear relations of student traits to levels of stress.

3.3.2 Decision Tree Classifier

The Decision Tree Classifier models the decision-making process as a hierarchical tree structure, where internal nodes represent feature-based decisions and leaf nodes represent predicted stress categories. This algorithm is particularly effective for capturing non-linear relationships and feature interactions that are common in survey-based stress data.

Decision Trees offer high interpretability, as predictions can be explained through explicit decision rules. In this implementation, class weighting is applied to mitigate class imbalance. However, Decision Trees are prone to overfitting, especially when trained on high-dimensional data produced by one-hot encoding. This limitation motivates the use of ensemble methods for improved generalisation.

3.3.3 Random Forest Classifier

The Random Forest Classifier is an ensemble learning method that combines multiple decision trees trained on different subsets of the data and feature space. Predictions are generated using majority voting across trees. This ensemble approach significantly reduces overfitting while improving robustness and generalisation.

In this solution, Random Forest models are trained using class weighting to ensure fair learning across all stress categories. Compared to single Decision Trees, Random Forest demonstrates stronger predictive performance and greater stability on unseen data. Additionally, Random Forest provides feature importance scores, which support interpretability by identifying survey attributes that contribute most strongly to stress predictions.

3.3.4 Hyperparameter-tuned Random Forest

Apart from the basic Random Forest algorithm, the tuned Random Forest algorithm is implemented in this project to boost the predictions. Parameters of the algorithm, which include the number of trees in the forest, the complexity of the tree, and the minimum samples required for a split and for a node, are tuned.

The Tuned Random Forest shows promising classification performance over the baseline models, especially when dealing with minority stress categories. This variant proves the value of model tuning for obtaining accurate results when processing high-dimensional survey data. This model later forms the basis for the final three-class ordinal stress prediction approach presented in Section 3.8.5

3.3.5 Summary of Algorithm Suitability

These algorithms were chosen to ensure interpretability, robustness, and good prediction performance. Logistic Regression allows for interpretability and a good starting performance, Decision Trees can model non-linear decision-making logic, and Random Forest can improve prediction performance due to the ensemble approach.

Algorithm	Key Characteristics	Role in This Solution
Logistic Regression	Linear, probabilistic, interpretable	Baseline comparison model
Decision Tree	Non-linear, rule-based, interpretable	Captures complex feature relationships
Random Forest	Ensemble-based, robust, reduced overfitting	Primary high-performance model
Tuned Random Forest	Optimised ensemble model	Best overall performance and stability

Table 3 Summary table of Implemented Algorithms

3.4 Pseudocode

```
START StudentStressPredictionSystem
   IMPORT pandas
IMPORT numpy


IMPORT scikit-learn:
LabelEncoder
train_test_split
LogisticRegression
DecisionTreeClassifier
RandomForestClassifier
evaluation metrics
(accuracy, precision, recall, F1-score, confusion matrix)


IMPORT matplotlib
IMPORT seaborn
                        Dataset exploration
   LOAD student stress survey dataset
   READ survey dataset into data frame
   DISPLAY dataset shape (rows and columns)
   DISPLAY column names and data types
   CHECK for missing values
```

```
IF missing values exist THEN
    HANDLE missing values appropriately
ELSE
    CONTINUE processing
```

Exploratory Data Analysis

```
ANALYSE distribution of stress levels
PLOT class distribution to identify imbalance

FOR selected academic features DO
    PLOT feature values against stress levels
END FOR

FOR selected health and lifestyle features DO
    PLOT feature values against stress levels
END FOR

INTERPRET patterns and relationships
DOCUMENT insights for feature relevance and modelling decisions
```

Data Preprocessing

```
REMOVE identifier columns not useful for prediction

SEPARATE features (X) and target variable (y)
```

```
ENCODE categorical features using one-hot encoding
ENCODE target stress labels using label encoding

SPLIT dataset into training set and test set
    USE stratified sampling to preserve class distribution
```

3.4.1 Pseudocode for Logistic Regression Model

```
INITIALISE Logistic Regression classifier
    SET multi-class strategy
    SET class_weight to "balanced"

TRAIN model using training data

PREDICT stress levels on test data

EVALUATE predictions
    CALCULATE accuracy
    CALCULATE precision, recall, F1-score
    GENERATE confusion matrix
```

3.4.2 Pseudocode for Decision Tree Model

```
INITIALISE Decision Tree classifier
    SET class_weight to "balanced"

TRAIN model using training data

PREDICT stress levels on test data

EVALUATE predictions
    CALCULATE accuracy
    CALCULATE precision, recall, F1-score
    GENERATE confusion matrix

NOTE potential overfitting behaviour
```

3.4.3 Pseudocode for Random Forest Model

```
INITIALISE Random Forest classifier
    SET number of trees
    SET random seed
    SET class_weight to "balanced"
```

```
TRAIN model using training data

PREDICT stress levels on test data

EVALUATE predictions
    CALCULATE accuracy
    CALCULATE precision, recall, F1-score
    GENERATE confusion matrix

EXTRACT feature importance values
```

3.4.4 Pseudocode for Tuned Random Forest Model

```
DEFINE hyperparameter search space
    Number of trees
    Tree depth
    Minimum samples for split
    Minimum samples for leaf

TRAIN Random Forest using tuned parameters

PREDICT stress levels on test data
```

```
    EVALUATE performance
        COMPARE results with baseline models

END TunedRandomForestModel
```

3.4.5 Pseudocode for Ordinal Random Forest Model

```
    DEFINE mapping function for stress levels
        MAP "Very Low" and "Low" → Low
        MAP "Moderate" → Moderate
        MAP "High" and "Very High" → High

    APPLY mapping function to target variable
    CREATE new ordinal stress label

    ENCODE ordinal stress labels numerically

    SELECT preprocessed features
    SELECT ordinal stress target variable

    SPLIT data into training and testing sets
        USE stratified sampling
```

```
INITIALISE Random Forest classifier
    SET number of trees
    SET class_weight to "balanced"

TRAIN model using training data

PREDICT ordinal stress levels on test data
PREDICT class probabilities

CALCULATE accuracy
GENERATE classification report
    Precision
    Recall
    F1-score
    Macro average
    Weighted average

GENERATE confusion matrix
    ANALYSE misclassification patterns

BINARISE true labels for multi-class evaluation

FOR each stress class DO
```

```
        COMPUTE ROC curve

        COMPUTE AUC score

        COMPUTE Precision-Recall curve

        COMPUTE Average Precision score

    END FOR

    INTERPRET evaluation results


    COMPARE baseline and ordinal model results

    ANALYSE impact of class imbalance handling

    ANALYSE improvement after ordinal reformulation


    SELECT ordinal Random Forest as final model

        BASED ON performance

        BASED ON interpretability

        BASED ON real-world suitability

END StudentStressPredictionApplication
```

3.5 State Transition Representation

Figure 4 contains the transition of the state of the implemented system-from the raw student survey data to predicted stress categories. Indeed, the system transitions a state: input (raw survey responses), which involves processing that includes cleaning and pre-processing features, encoding features, splitting datasets, model training and evaluation, and model inference; output-low, moderate, or high decoded and interpretable classification, useful for supporting early identification of students with high levels of perceived stress.

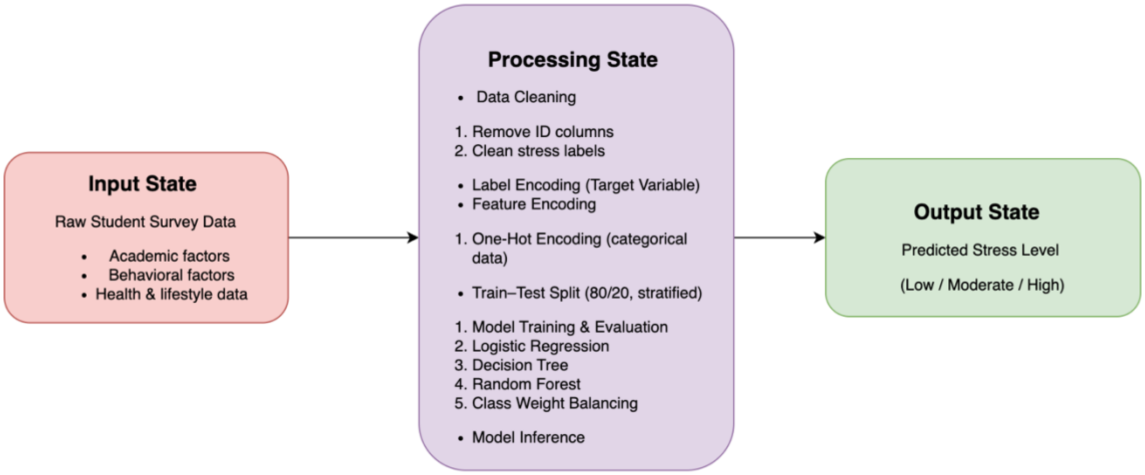


Figure 3 State Transition Representation

3.6 Flowchart Representation

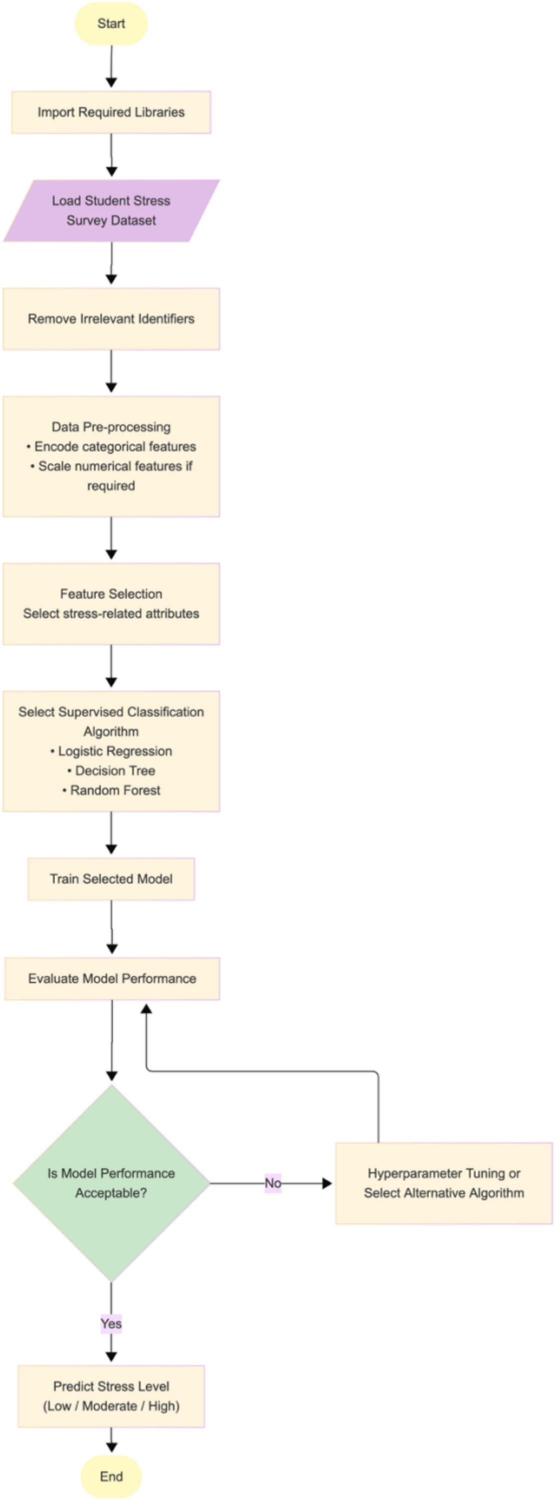

Start
Import Required Libraries
Load Student Stress Survey Dataset
Remove Irrelevant Identifiers
Data Pre-processing
• Encode categorical features
• Scale numerical features if required
Feature Selection
Select stress-related attributes
Select Supervised Classification Algorithm
• Logistic Regression
• Decision Tree
• Random Forest
Train Selected Model
Evaluate Model Performance
Is Model Performance Acceptable?
No
Hyperparameter Tuning or Select Alternative Algorithm
Yes
Predict Stress Level (Low / Moderate / High)
End

Figure 4 Flowchart of overall system

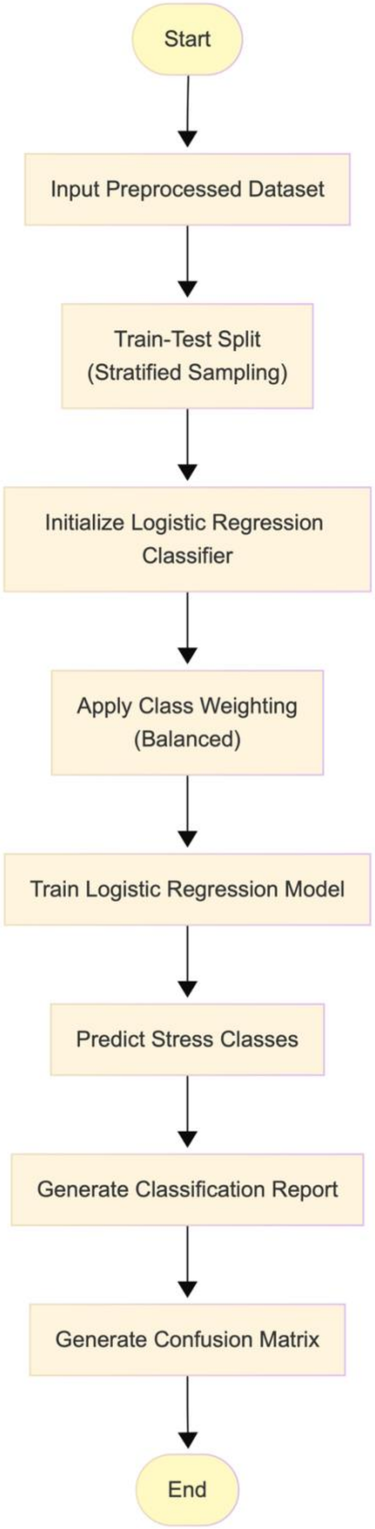

Start
Input Preprocessed Dataset
Train-Test Split
(Stratified Sampling)
Initialize Logistic Regression
Classifier
Apply Class Weighting
(Balanced)
Train Logistic Regression Model
Predict Stress Classes
Generate Classification Report
Generate Confusion Matrix
End

Figure 5 Logistic Regrssion Flowchart

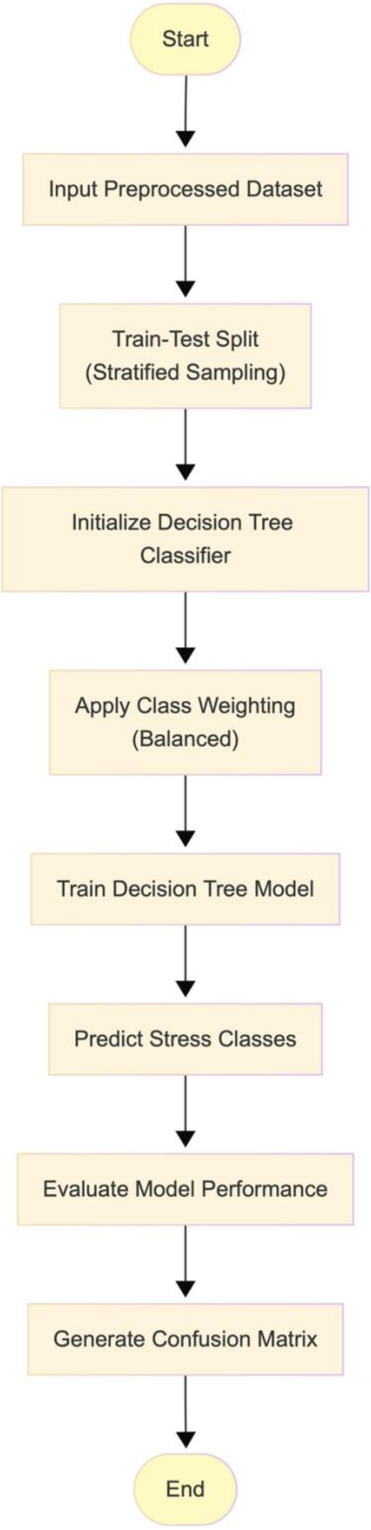


Figure 6 Decision Tree Flowchart

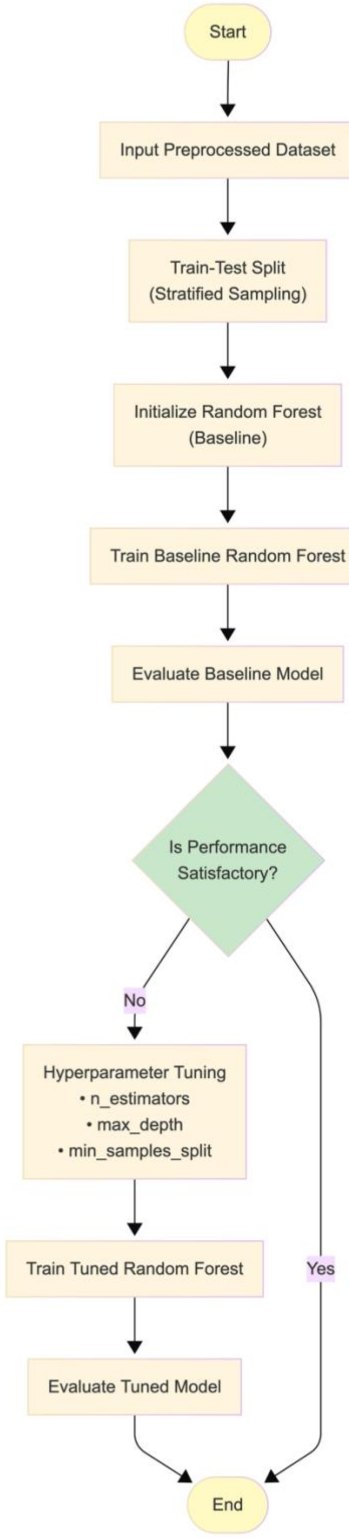


Figure 7 Random Forest (Baseline and Hyperparameter tuned) Flowchart

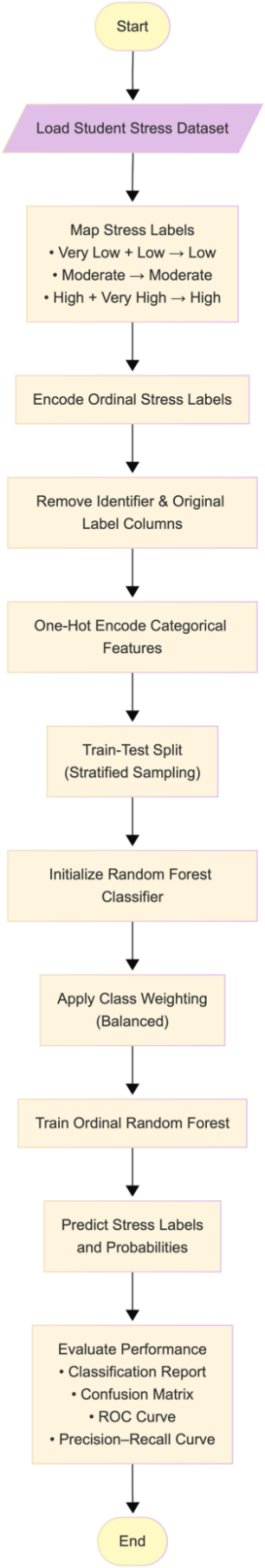

Start
Load Student Stress Dataset
Map Stress Labels
• Very Low + Low → Low
• Moderate → Moderate
• High + Very High → High
Encode Ordinal Stress Labels
Remove Identifier & Original Label Columns
One-Hot Encode Categorical Features
Train-Test Split (Stratified Sampling)
Initialize Random Forest Classifier
Apply Class Weighting (Balanced)
Train Ordinal Random Forest
Predict Stress Labels and Probabilities
Evaluate Performance
• Classification Report
• Confusion Matrix
• ROC Curve
• Precision–Recall Curve
End

Figure 8 Ordinal Random Forest Flowchart

3.7 Development Tools, Technologies, and Libraries Used

The development of the student stress classification system was carried out using Python and a set of widely adopted data science and machine learning tools. These tools supported data preprocessing, model implementation, evaluation, visualisation, and application deployment.

Category	Tool / Library	Purpose in the Solution
Programming Language	Python	Core language used to implement data preprocessing, machine learning models, evaluation, and prediction
Development Environment	Jupyter Notebook	Used for iterative development, experimentation, debugging, and visual inspection of model outputs
Code Editor	Visual Studio Code	Used for structured coding, file management, and application development
Data Manipulation	Pandas	Loading datasets, cleaning survey data, handling missing or inconsistent values, and preparing feature matrices
Numerical Computing	NumPy	Supporting numerical operations and array-based computations
Machine Learning	Scikit-learn	Implementation of Logistic Regression, Decision Tree, and Random Forest classifiers, model training, evaluation, and class balancing
Data Preprocessing	Scikit-learn (LabelEncoder)	Encoding categorical stress labels into numerical form
Feature Encoding	Pandas (get_dummies)	One-hot encoding of categorical survey features

Model Evaluation	Scikit-learn (metrics)	Generating classification reports, confusion matrices, accuracy, precision, recall, and F1-scores
Data Visualisation	Matplotlib, Seaborn	Visualisation of confusion matrices and model performance results
Web Application Framework	Streamlit	Development of an interactive interface to demonstrate stress-level prediction for new inputs
Version Control	Git & GitHub	Source code versioning, progress tracking, and backup of project files

Table 4 Development Tools and Technologies

3.8 Achieved Results

This section presents some EDA on the dataset, shows the class distribution and imbalance analysis along with the results obtained from implementing multiple supervised machine learning models for student stress classification. The evaluation follows a progressive experimentation strategy, beginning with baseline multi-class classification and advancing towards an ordinal reformulation of the problem. Model performance is analysed using accuracy, precision, recall, F1-score, and confusion matrices, with particular attention given to class imbalance and class overlap, which significantly influenced results.

3.8.1 Class Distribution and Imbalance Analysis

Initial analysis of the dataset revealed a highly imbalanced target variable, particularly in the original multi-class formulation of stress levels. Some stress categories contained substantially fewer samples than others, leading to biased learning and unstable predictions. Minority classes suffered from low recall, meaning stressed students were often not correctly identified.

Although class weighting was applied during training to mitigate this issue, imbalance combined with overlapping survey features made accurate separation of all five stress classes challenging. This imbalance played a key role in the performance limitations observed in early experiments.

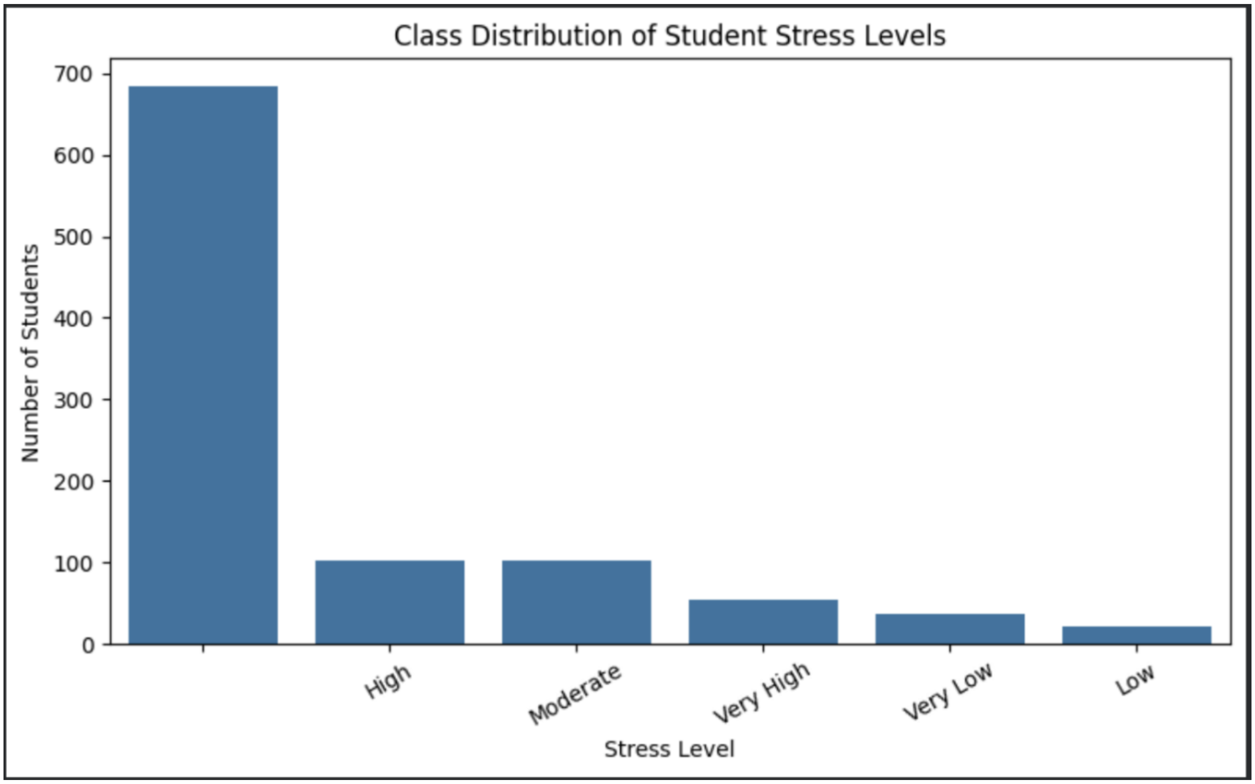


Figure 9 Class Imbalance in the dataset

3.8.2 Exploratory Data Analysis

Following the class distribution analysis, further exploratory data analysis was conducted to examine the relationships between key academic, behavioural, lifestyle, and health-related factors and student stress levels. These analyses aim to validate feature relevance and provide insight into the multi-dimensional nature of stress among students.

Analysis of academic workload revealed a clear association between perceived workload intensity and reported stress levels. Students who indicated higher academic workload contribution were more frequently observed in moderate to high stress categories, highlighting academic pressure as a significant contributor to student stress.

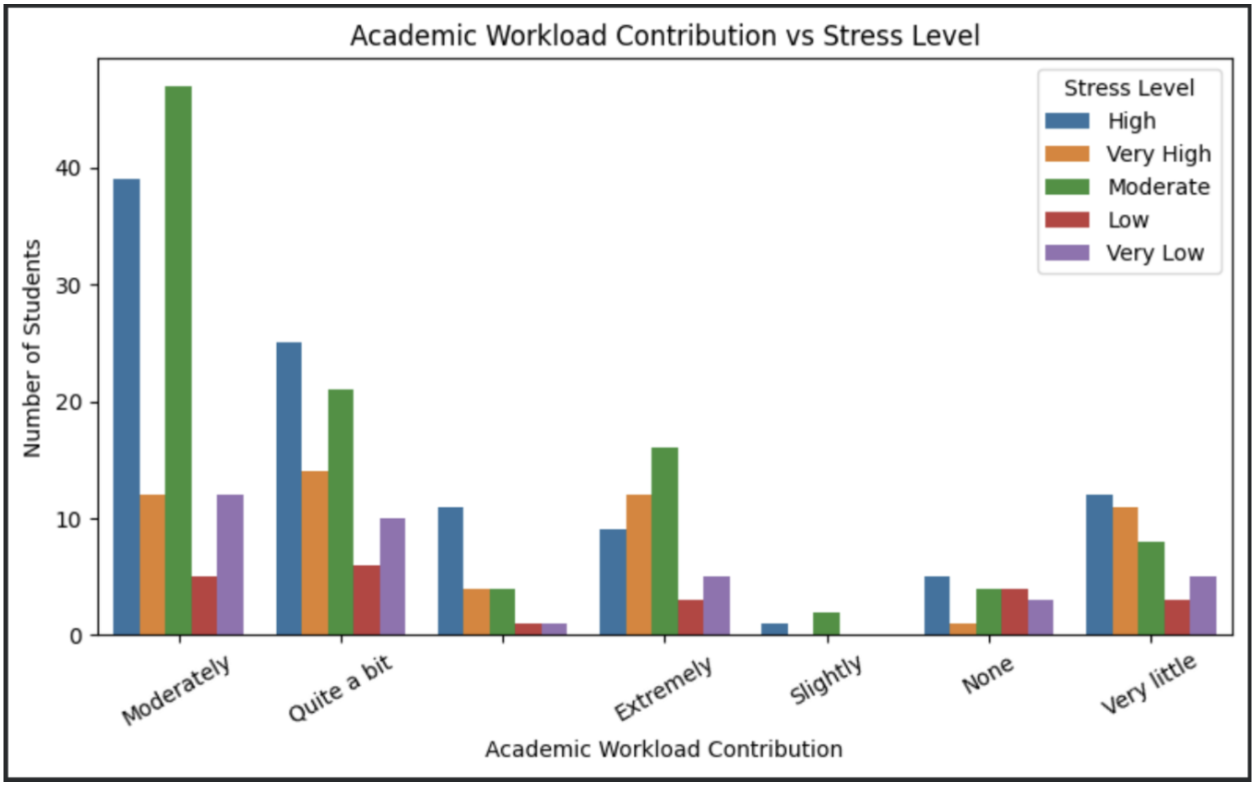


Figure 10 Academic Workload Contribution v/s Stress

Daily study duration was also examined to understand behavioural patterns related to stress. The results suggest that students reporting longer study hours per day tend to exhibit higher stress levels, particularly when combined with other academic demands. This supports the inclusion of study behaviour variables in stress prediction models.

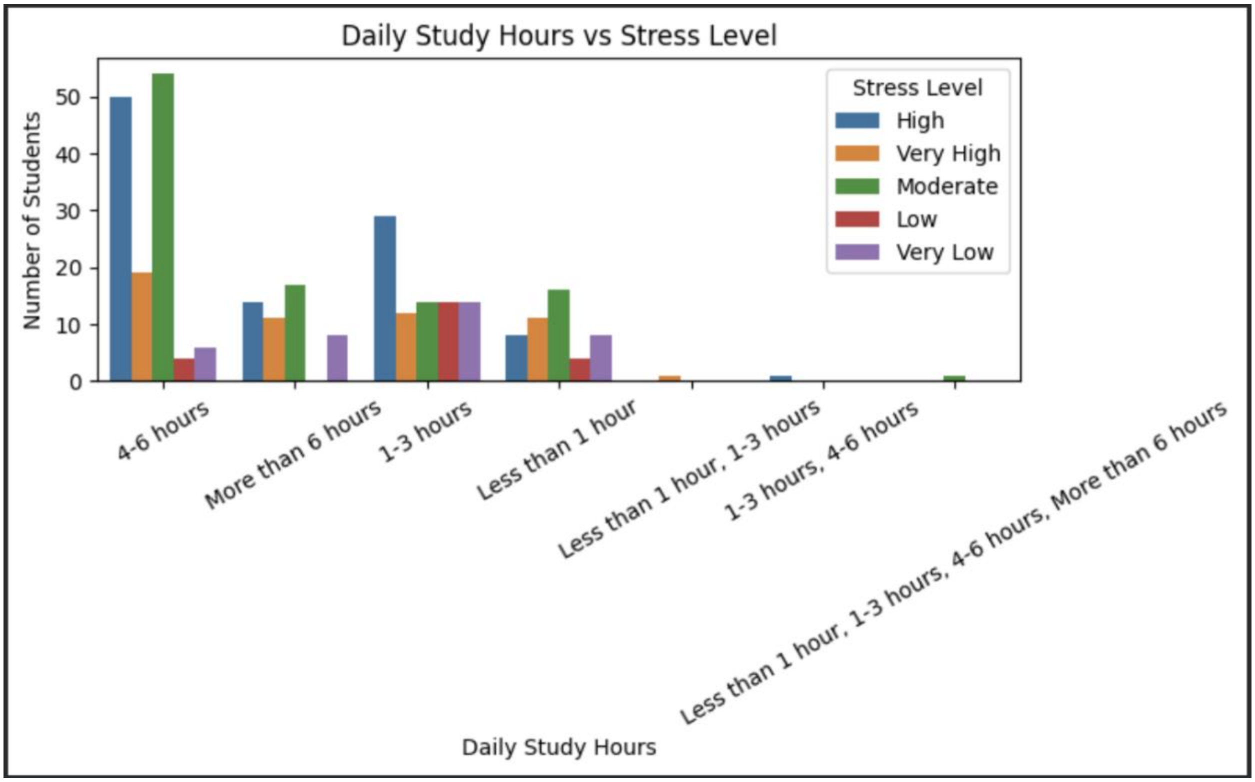


Figure 11 Daily Study Hours v/s Stress

There was a strong relationship between the level of stress experienced by students and their sleep quality. Students with poor or irregular sleep quality were found to have been mostly categorized under higher levels of stress, while those with good sleep quality were mostly found to have been under lower levels of stress. However, this is not surprising due to associations seen in existing research.

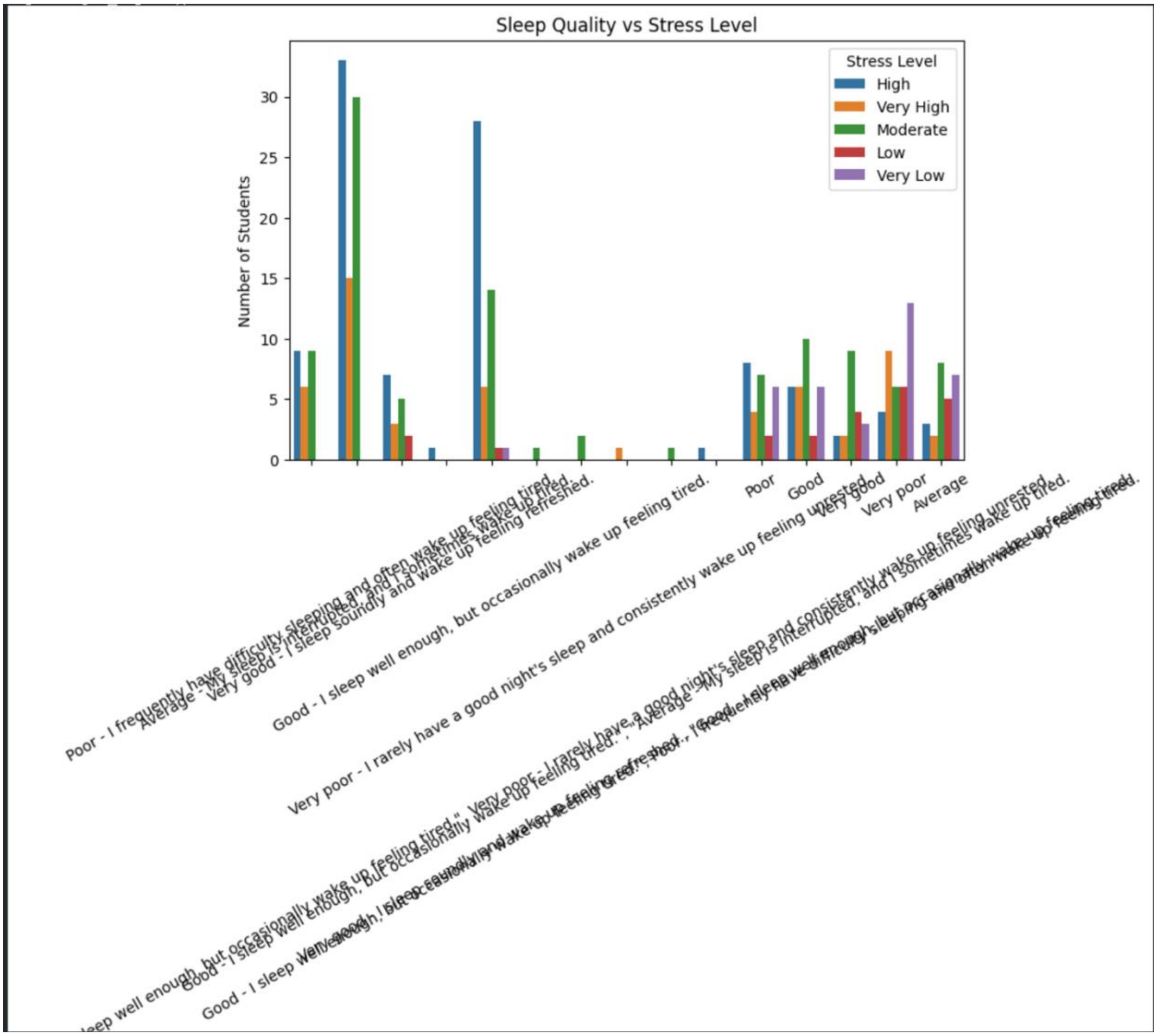


Figure 12 Sleep Quality v/s Stress

Further observations were made on the pressure associated with examination work. Students who strongly agreed on the pressure of examination performance influencing their stress largely tended to be categorized in high stress categories. This validates the examination environment as an influential source of stress in the learning environment.

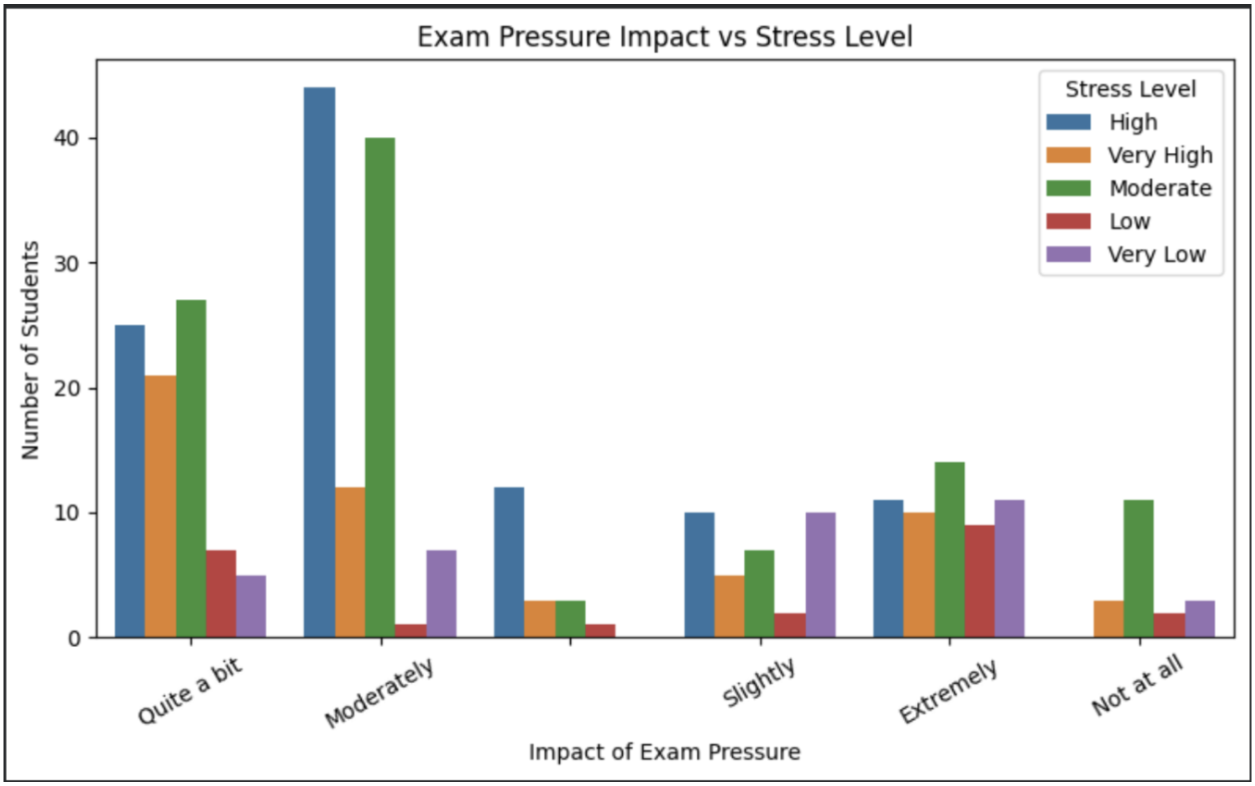


Figure 13 Exam Pressure Impact v/s Stress

Finally, the analysis turned to extra-curricular involvement to see if there was any balancing effect being made. It would appear from the data that the groups of extra-curricular involved students are more randomly dispersed throughout the stress levels than the non-extra-curricular involved students.

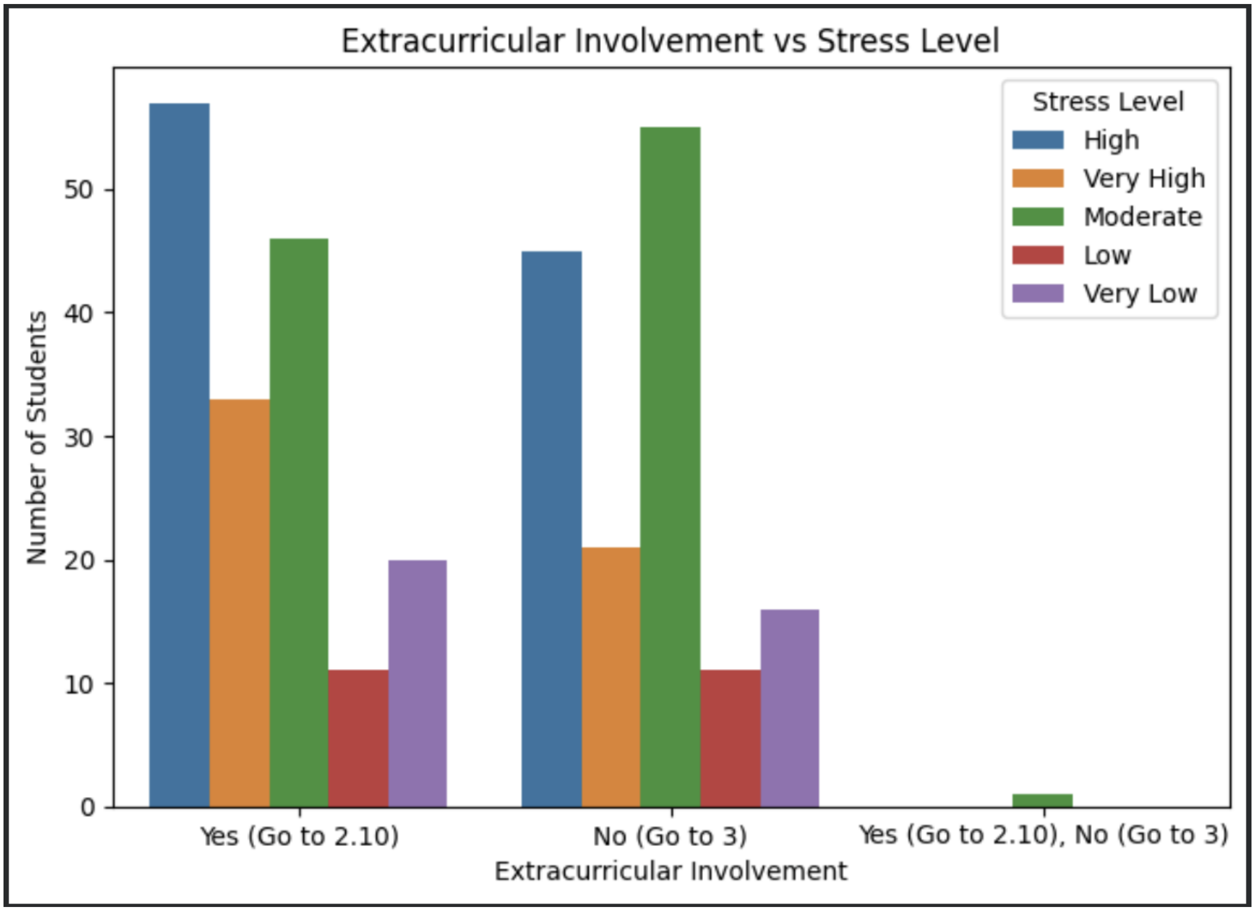


Figure 14 Extracurricular Involvement v/s Stress

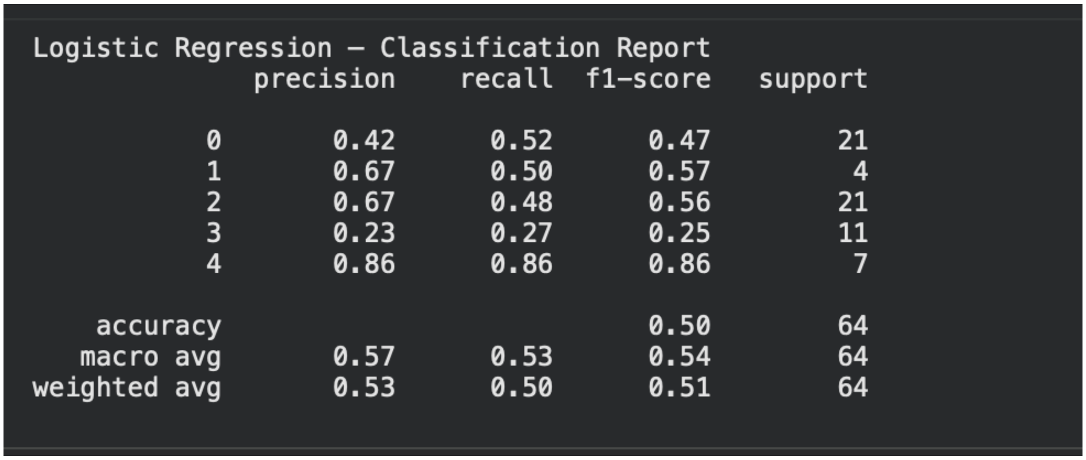
On the whole, the exploratory data analysis has proved that student stress is a phenomenon influenced by various factors, including academic workload, study behavior, quality of sleep, examination pressure, and lifestyle factors. This validates the fact that multiple feature supervised learning models can and should have been employed for analysis and problem formulation.

3.8.3 Multi-Class (Five-Class) Classification Results

Baseline: Logistic Regression

Logistic Regression was implemented as the baseline classifier. The model achieved an accuracy of 50%, with a macro F1-score of 0.54. While it performed reasonably for majority classes, it struggled significantly with minority and mid-range stress categories.

The confusion matrix revealed frequent misclassification between adjacent stress levels, particularly among moderate stress classes. This behaviour indicates that the linear decision boundaries of Logistic Regression were insufficient to capture the complex, non-linear interactions present in survey-based stress data.



```
Logistic Regression — Classification Report
              precision    recall  f1-score   support

           0       0.42      0.52      0.47        21
           1       0.67      0.50      0.57         4
           2       0.67      0.48      0.56        21
           3       0.23      0.27      0.25        11
           4       0.86      0.86      0.86         7

    accuracy                           0.50        64
   macro avg       0.57      0.53      0.54        64
weighted avg       0.53      0.50      0.51        64
```

Figure 15 Logistic Regression Classification Report

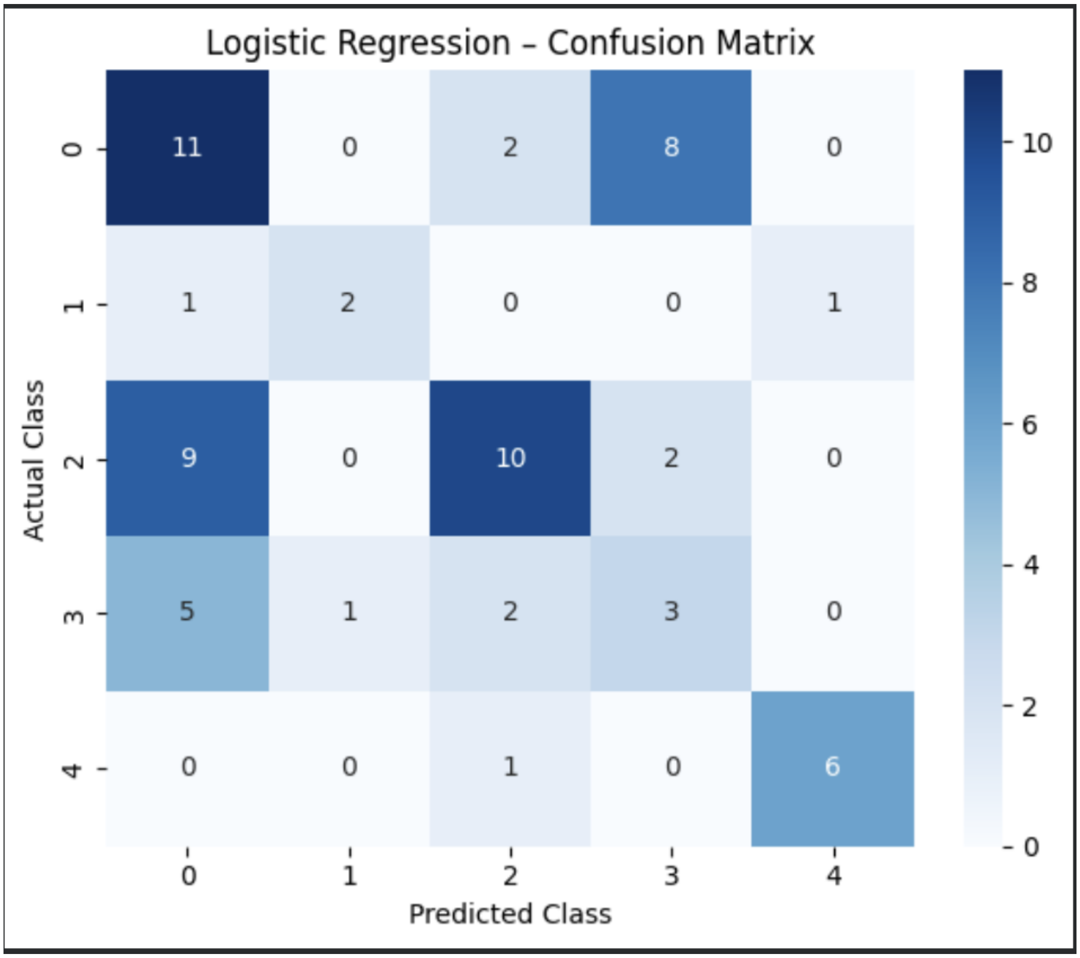
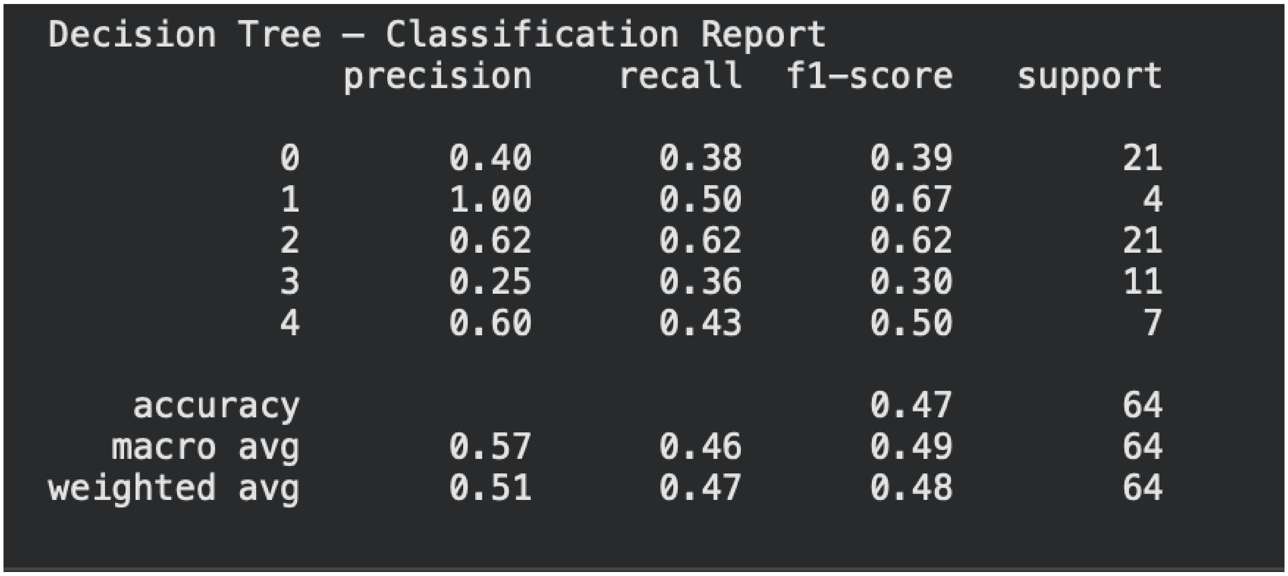


Figure 16 Logistic Regression Confusion Matrix

Decision Tree Classifier

The Decision Tree classifier was introduced to model non-linear feature interactions. However, despite improved interpretability, overall performance declined slightly (accuracy = 46.9%). The model showed sensitivity to noise and overfitting, particularly due to the high dimensionality introduced by one-hot encoding.

Although some classes achieved high precision, recall remained inconsistent, especially for minority stress categories. This demonstrated that a single tree lacked the robustness required for generalisation.



```
Decision Tree — Classification Report
              precision    recall  f1-score   support

           0       0.40      0.38      0.39        21
           1       1.00      0.50      0.67         4
           2       0.62      0.62      0.62        21
           3       0.25      0.36      0.30        11
           4       0.60      0.43      0.50         7

    accuracy                           0.47        64
   macro avg       0.57      0.46      0.49        64
weighted avg       0.51      0.47      0.48        64
```

Figure 17 Decision Tree Classification Report

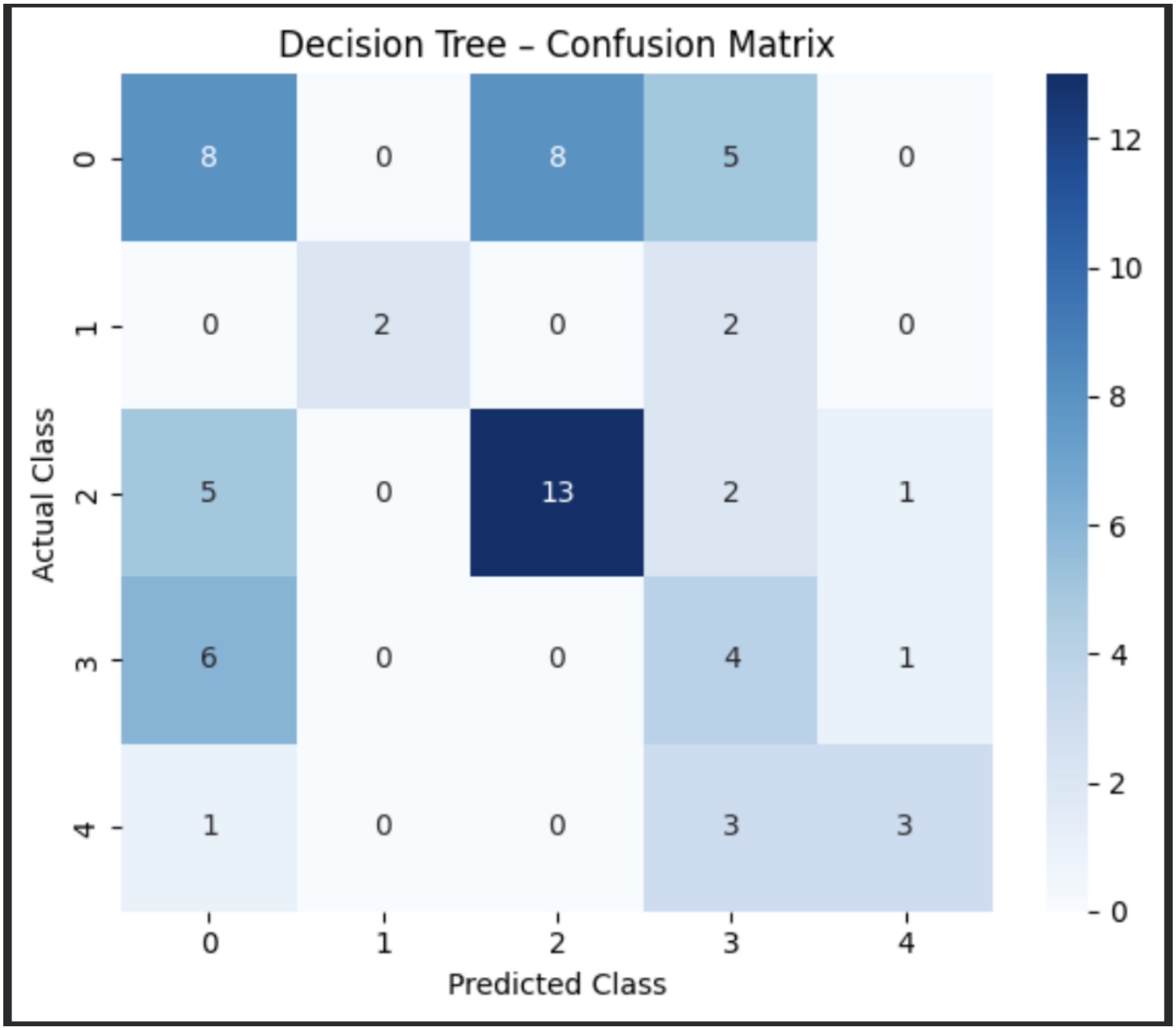
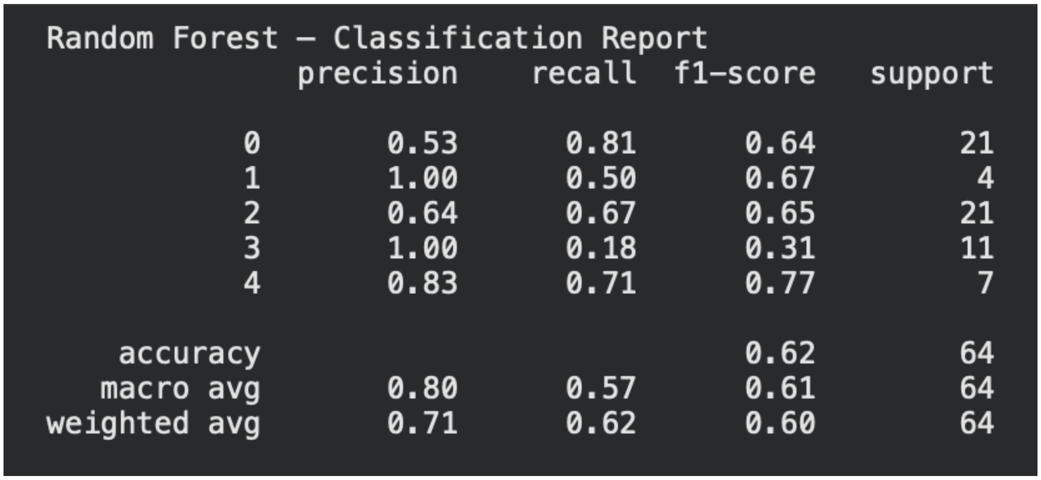


Figure 18 Decision Tree Confusion Matrix

Random Forest Classifier

The Random Forest classifier significantly improved performance, achieving an accuracy of 62% and a macro F1-score of 0.61. Ensemble learning reduced overfitting and improved generalisation across most stress categories.

The confusion matrix showed a stronger diagonal concentration, indicating more consistent predictions. However, despite this improvement, mid-range stress categories remained difficult to distinguish, with persistent confusion between adjacent classes. This highlighted a fundamental issue: the five-class formulation introduced artificial boundaries between stress levels that were semantically similar.



```
Random Forest – Classification Report
              precision    recall  f1-score   support

           0       0.53      0.81      0.64        21
           1       1.00      0.50      0.67         4
           2       0.64      0.67      0.65        21
           3       1.00      0.18      0.31        11
           4       0.83      0.71      0.77         7

    accuracy                           0.62        64
   macro avg       0.80      0.57      0.61        64
weighted avg       0.71      0.62      0.60        64
```

Figure 19 Random Forest Classification Report

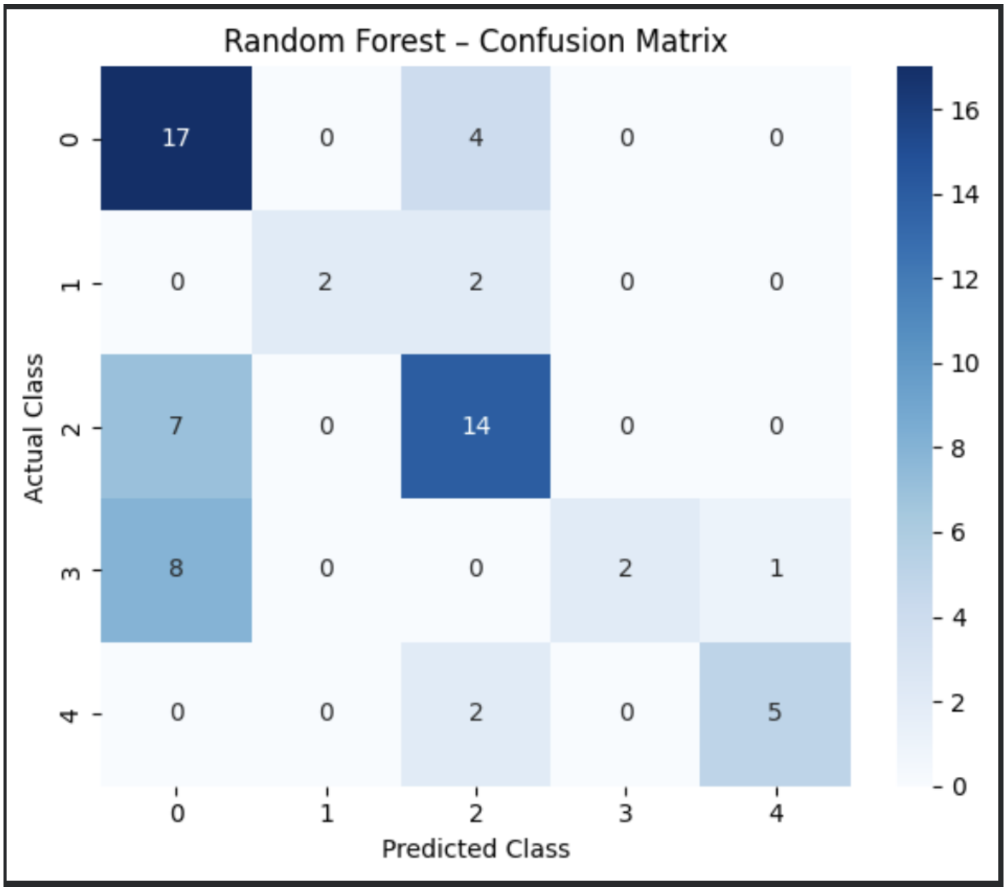
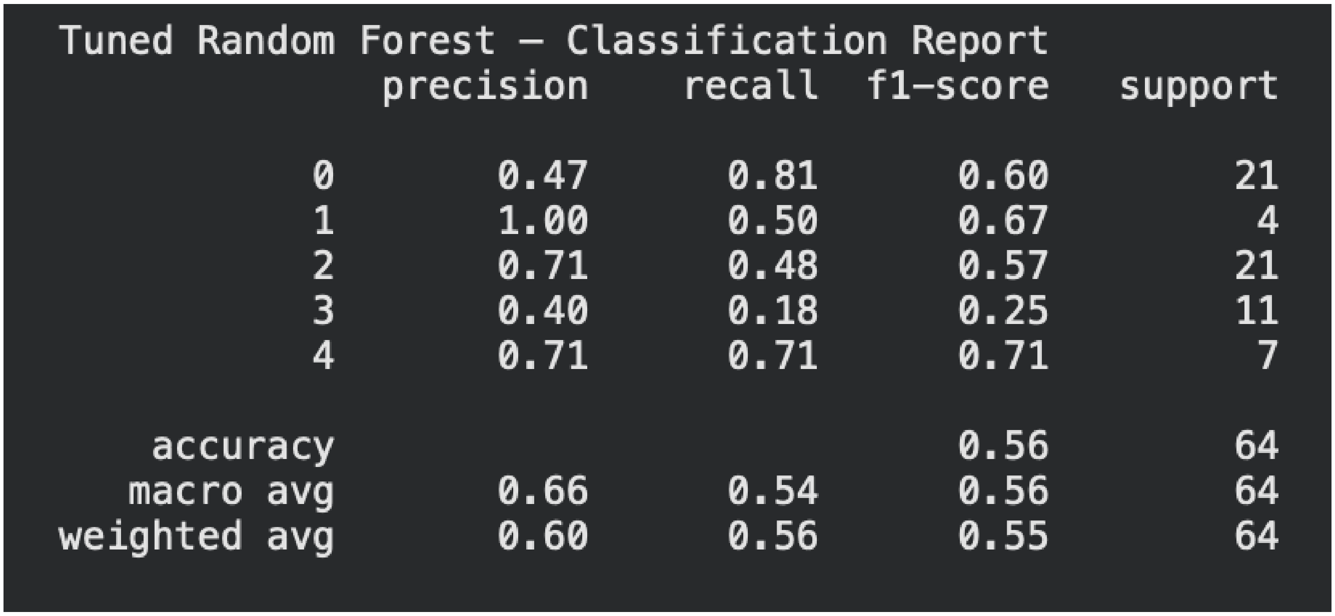


Figure 20 Random Forest Confusion Matrix

Tuned Random Forest

Hyperparameter tuning was applied to further optimise the Random Forest model. Contrary to expectations, tuning resulted in reduced performance (accuracy = 56%). The tuned model showed signs of underfitting, particularly for moderate stress classes, suggesting that excessive constraint on tree complexity limited the model's expressive power.

This outcome demonstrates that hyperparameter tuning does not always guarantee improvement, especially when data size is limited and class boundaries are inherently ambiguous.



```
Tuned Random Forest — Classification Report
              precision    recall  f1-score   support

           0       0.47      0.81      0.60        21
           1       1.00      0.50      0.67         4
           2       0.71      0.48      0.57        21
           3       0.40      0.18      0.25        11
           4       0.71      0.71      0.71         7

    accuracy                           0.56        64
   macro avg       0.66      0.54      0.56        64
weighted avg       0.60      0.56      0.55        64
```

Figure 21 Tuned Random Forest Classification Report

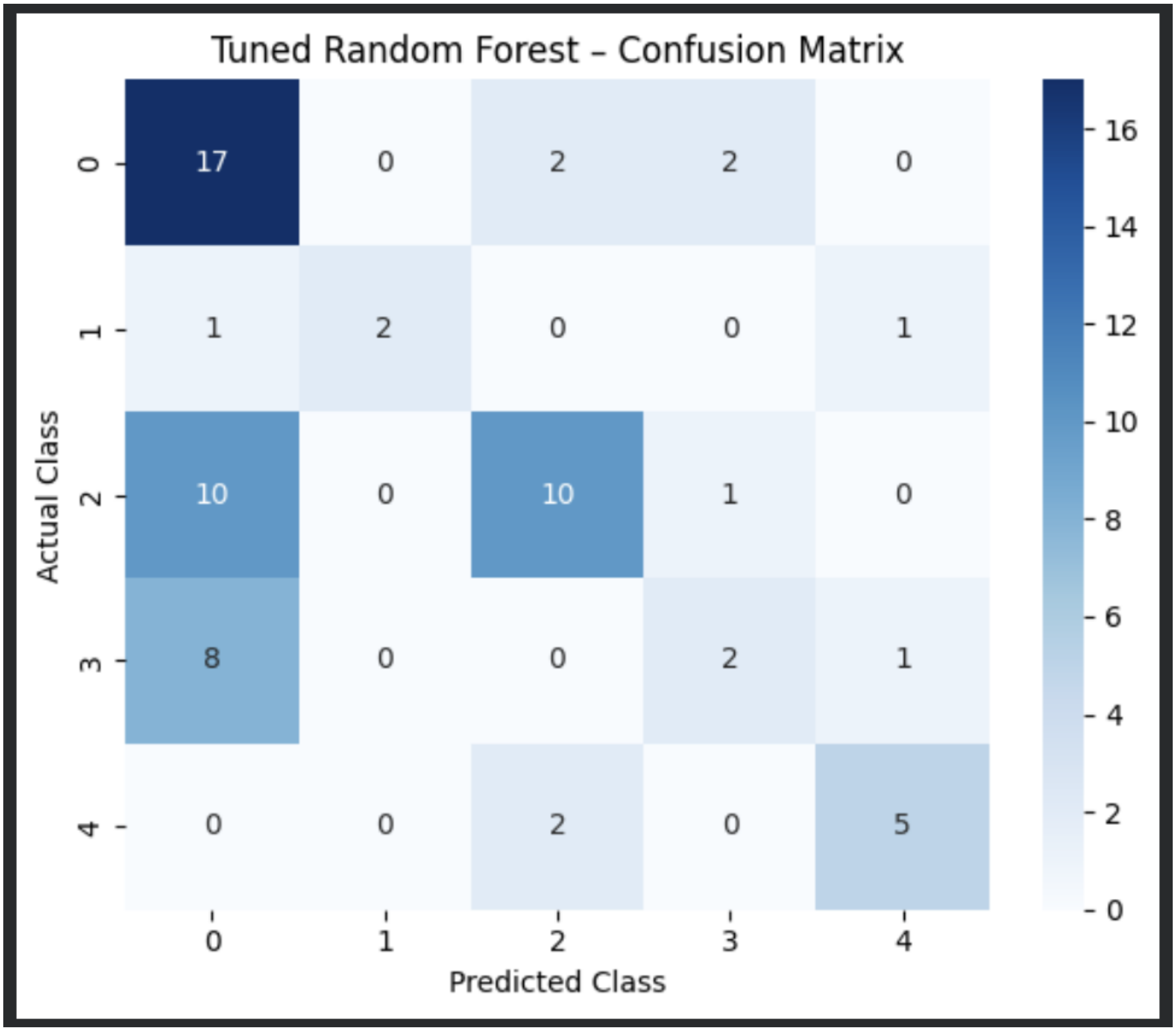


Figure 22 Tuned Random Forest Confusion Matrix

3.8.4 Motivation for Reformulating the Problem (Five-Class → Three-Class)

Analysis of the multi-class results revealed two key issues:

- Severe class overlap between adjacent stress levels
- Unstable learning for minority classes despite class weighting

From a domain perspective, stress severity is naturally ordinal, not categorical. Predicting fine-grained distinctions (e.g., “Very Low” vs “Low”) is both difficult and less actionable in real educational contexts.

To address this, the problem was reformulated into three ordered stress categories:

- Low
- Moderate
- High

This reduced class fragmentation, improved balance, and aligned better with real-world decision-making requirements.

3.8.5 Ordinal Random Forest Results (Three-Class Classification)

After reformulating the target variable, a Random Forest classifier was trained using the same preprocessing pipeline. This approach produced the best results across all experiments.

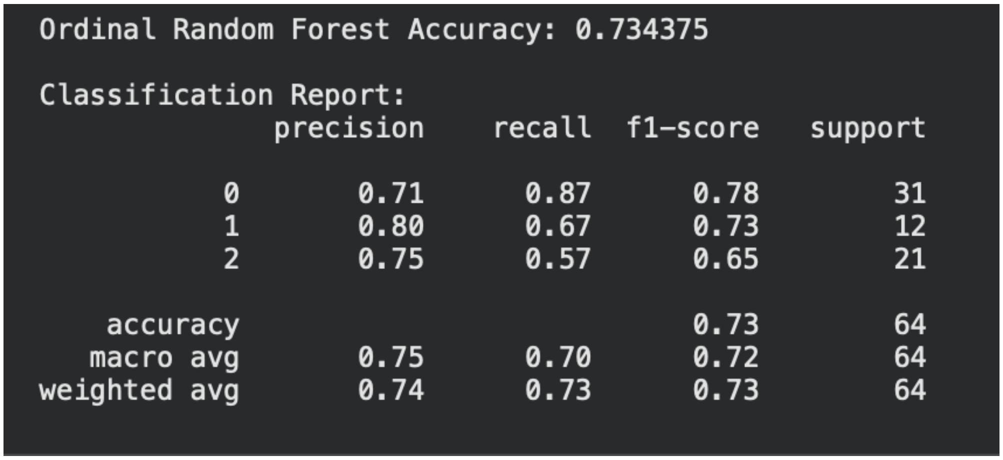
Performance Metrics (Ordinal Random Forest):

- Accuracy: 73%
- Macro F1-score: 0.72
- Weighted F1-score: 0.73

Class-level analysis showed strong and balanced performance:

- Low stress: Recall = 0.87, F1-score = 0.78
- Moderate stress: Balanced precision and recall (F1 = 0.73)
- High stress: Reliable identification (F1 = 0.65)

The confusion matrix demonstrated minimal severe misclassifications, with most errors occurring between adjacent ordinal classes, which is expected and acceptable in stress modelling.



```
Ordinal Random Forest Accuracy: 0.734375

Classification Report:
              precision    recall  f1-score   support

           0       0.71      0.87      0.78        31
           1       0.80      0.67      0.73        12
           2       0.75      0.57      0.65        21

    accuracy                           0.73        64
   macro avg       0.75      0.70      0.72        64
weighted avg       0.74      0.73      0.73        64
```

Figure 23 Ordinal Random Forest Classification Report

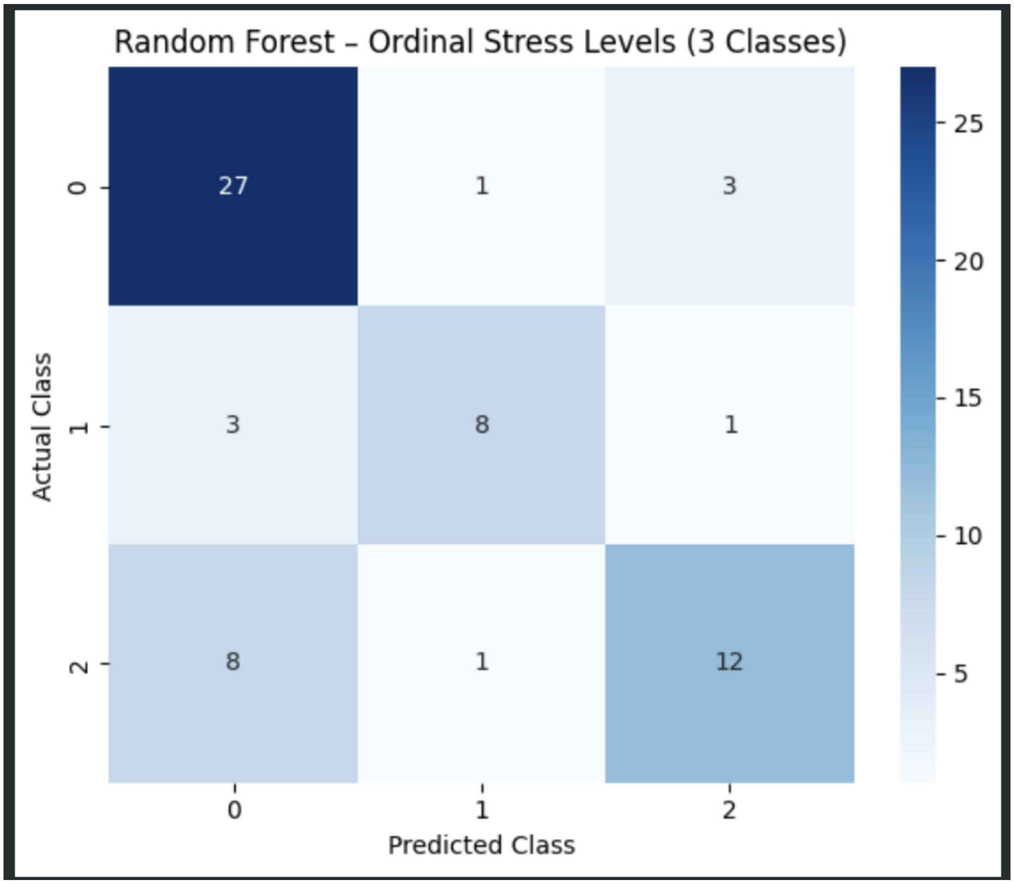


Figure 24 Ordinal Random Forest Confusion Matrix

ROC and Precision-Recall Analysis

Figure 25 shows the One-vs-Rest ROC curves of the ordinal Random Forest classifier. All stress categories have high AUC values, but Low stress has outstanding discriminative abilities (AUC of 0.96) compared to High stress (AUC of 0.88) and Moderate stress (AUC of 0.83). These findings suggest that there is a clear delineation among the stress categories, and particularly the ends of the ordinal range, that the model is capable of distinguishing. It is not surprising that there is a slightly lower AUC value in Moderate stress, as it lies in the middle of Low and High stress categories, yielding a larger overlap.

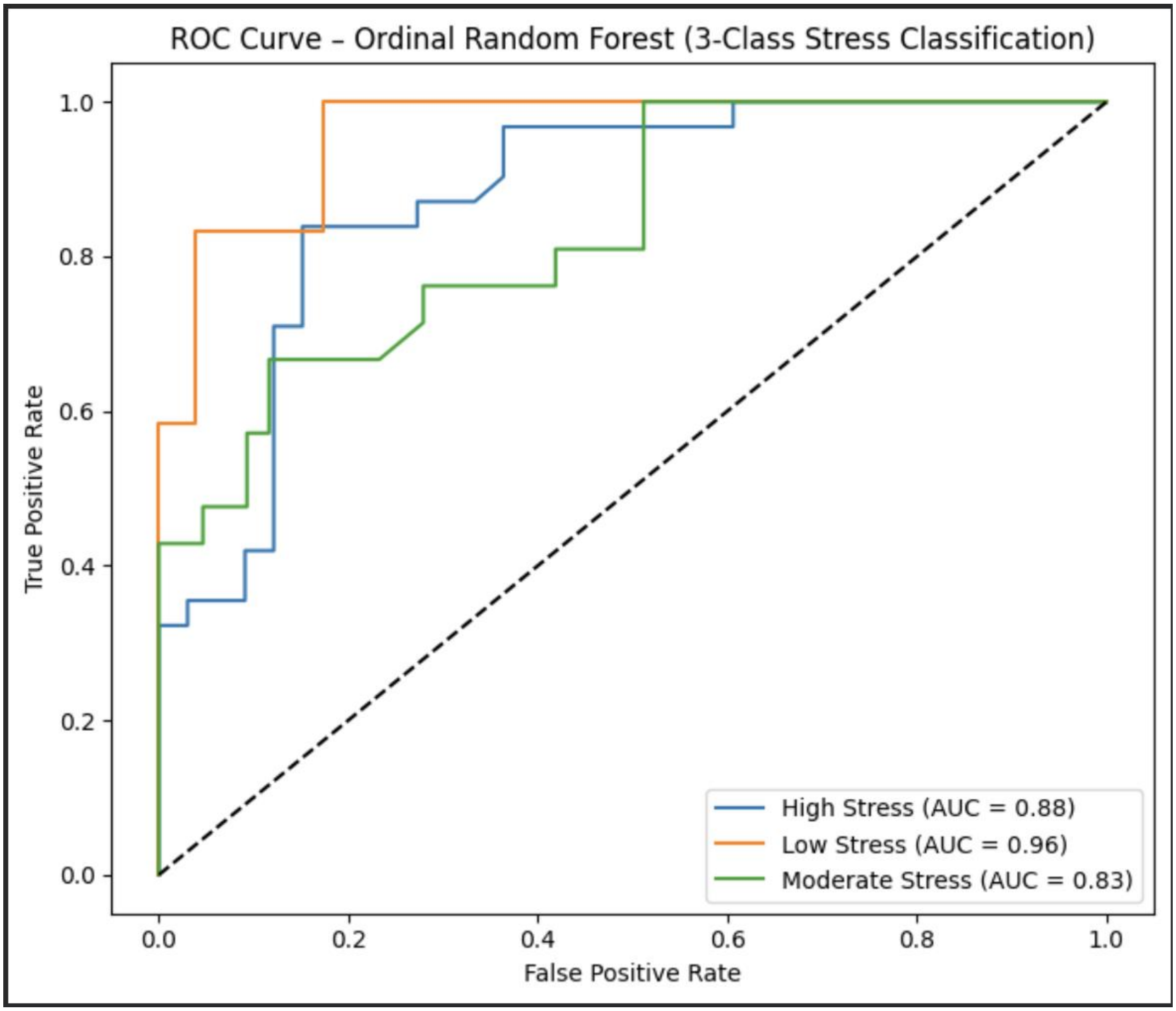


Figure 25 ROC Curve

Figure 26 above shows the Precision-Recall Curves of the same model to effectively convey the information in consideration of the class imbalance. The Average Precision scores in the table above further attest to the high performance of the model on Low stress (AP = 0.88) and High stress (AP = 0.86), as well as the acceptable performance on Moderate stress with an AP = 0.77, indicating high confidence in the model predictions of the former two categories of stress and the level of ambiguity of moderate stress levels as conveyed through the ordinal scale of stress severity.

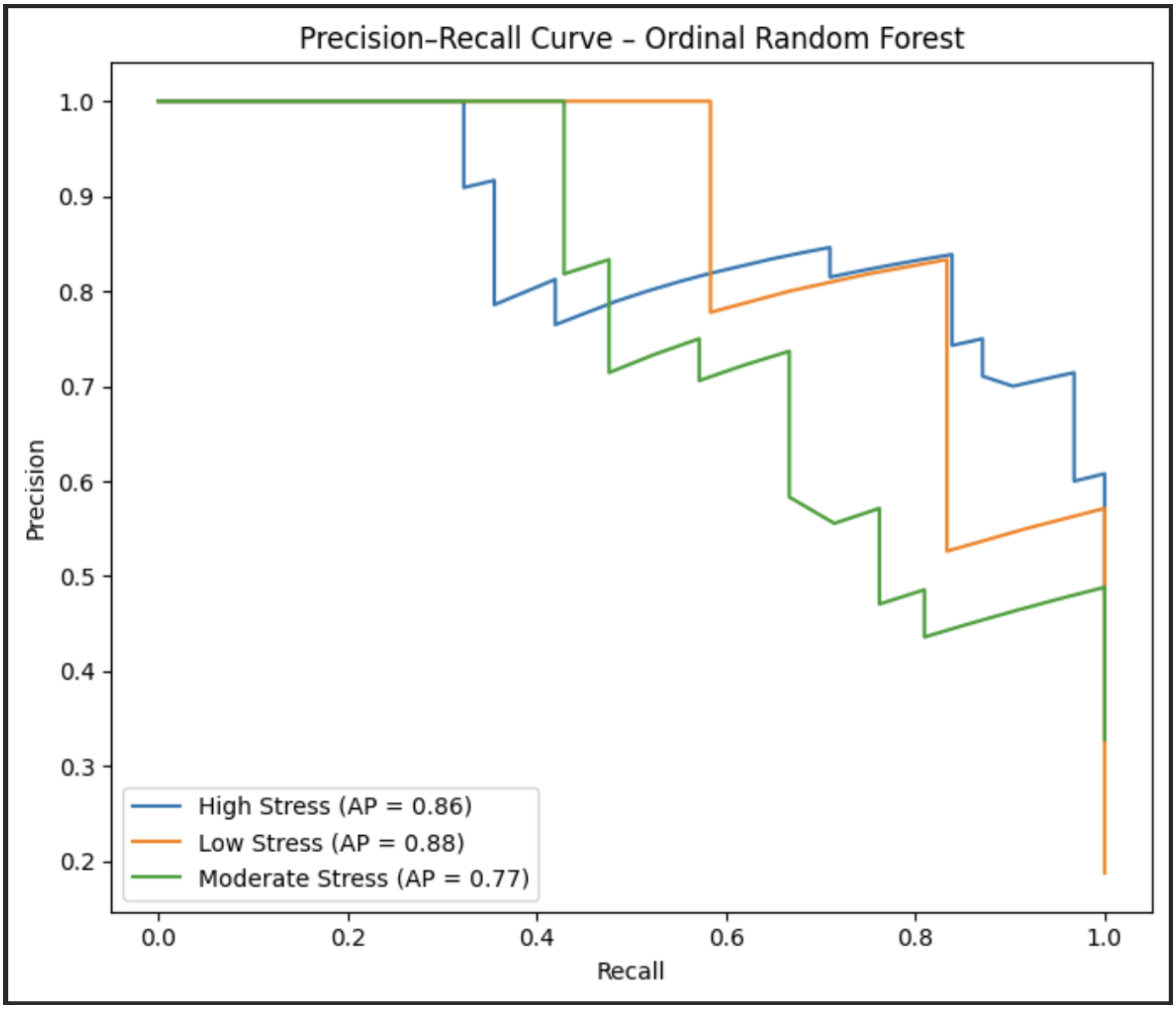


Figure 26 Precision Recall Curve

3.8.6 Comparative Model Performance Summary

Model	Accuracy	Macro F1	Key Observation
Logistic Regression (5-class)	0.50	0.54	Linear baseline, limited expressiveness
Decision Tree (5-class)	0.47	0.49	Non-linear but unstable
Random Forest (5-class)	0.63	0.61	Strong ensemble performance
Tuned Random Forest (5-class)	0.56	0.56	Underfitting after tuning
Ordinal Random Forest (3-class)	0.73	0.72	Best overall performance

Table 5 Comparative Model Performance Table

3.8.7 Key Findings and Interpretation

The achieved results clearly demonstrate that problem formulation is as important as model selection. While ensemble learning improved performance in the multi-class setting, the most significant gain was achieved by reformulating stress prediction as an ordinal classification task.

The ordinal Random Forest model provided:

- Improved accuracy and stability
- Better handling of class imbalance
- More interpretable and actionable outputs

This confirms that aligning machine learning design with domain semantics leads to more reliable and meaningful results.

3.8.8 Summary of Achieved Results

In summary, supervised machine learning techniques were successfully applied to classify student stress levels using structured survey data. Performance improved progressively through algorithm selection, ensemble learning, and ultimately problem reformulation. The ordinal Random Forest classifier achieved the strongest overall performance and represents the most suitable solution for real-world educational stress monitoring and early intervention.

3.9 Application

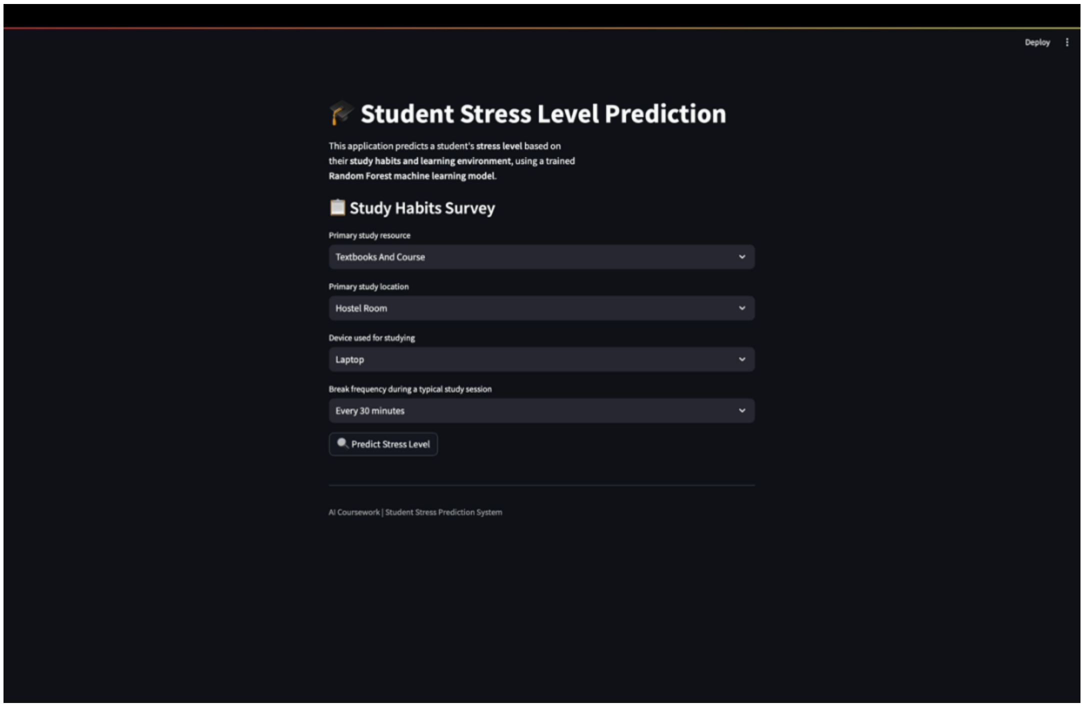
A simple Streamlit-based application was developed to demonstrate the practical use of the trained stress classification model. The application allows users to input student survey responses and receive a predicted stress level.

The application uses the final selected Random Forest model and applies the same preprocessing steps used during training to ensure consistency between model evaluation and prediction.

The workflow of the application is as follows:

- User enters student survey inputs.
- Input data is preprocessed and encoded.
- The trained model predicts the stress level.
- The predicted stress category (Low / Moderate / High) is displayed.

This application serves as a proof of concept to show how the proposed machine learning solution can be applied in a real educational context. It is intended for demonstration purposes only and not as a production-ready system.



🎓 Student Stress Level Prediction

This application predicts a student's **stress level** based on their **study habits and learning environment**, using a trained **Random Forest machine learning model.**

📋 Study Habits Survey

Primary study resource: Textbooks And Course

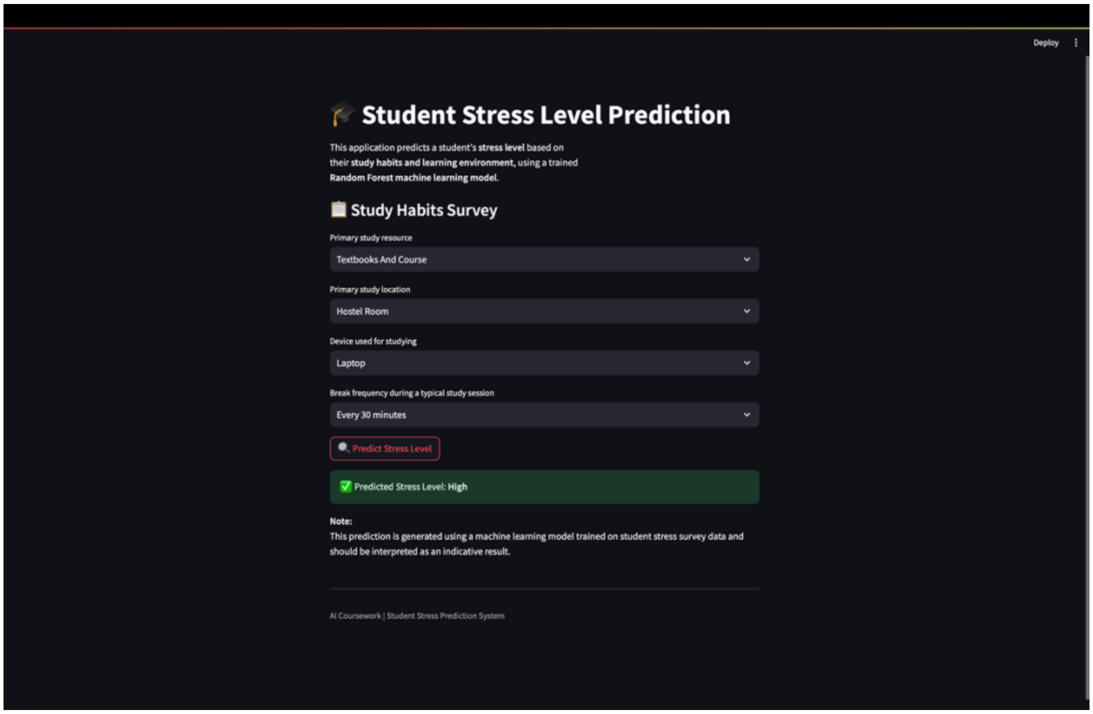
Primary study location: Hostel Room

Device used for studying: Laptop

Break frequency during a typical study session: Every 30 minutes

AI Coursework | Student Stress Prediction System

Figure 27 Application Interface



🎓 Student Stress Level Prediction

This application predicts a student's **stress level** based on their **study habits and learning environment**, using a trained **Random Forest machine learning model.**

📋 Study Habits Survey

Primary study resource: Textbooks And Course

Primary study location: Hostel Room

Device used for studying: Laptop

Break frequency during a typical study session: Every 30 minutes

✅ Predicted Stress Level: **High**

Note:
This prediction is generated using a machine learning model trained on student stress survey data and should be interpreted as an indicative result.

AI Coursework | Student Stress Prediction System

Figure 28 Application Output

4. Conclusion

4.1 Analysis of the Work Done

This coursework successfully demonstrated the application of supervised machine learning on the problem of student stress classification using structured survey data. Multiple classification models were explored in this study, like Logistic Regression, Decision Tree, and Random Forest, following a systematic machine learning pipeline comprising the steps of data preprocessing, feature encoding, training a model, model evaluation, and making predictions.

Initial experimentation using a five-class stress classification approach highlighted key challenges related to real-world educational data: class imbalance and class-overlap between adjacent classes of stress. We present how increasing classes will further exacerbate these issues, while class weighting and ensemble learning techniques somewhat improve the performance of a classifier, demonstrating that fine-grained multi-class stress prediction is an intrinsically hard problem to solve given the subjective and continuous nature of stress.

Through ensemble learning, the Random Forest classifier succeeded in effectively capturing non-linear relationships and remarkably reduced overfitting compared to baseline models. The most significant improvement came from reformulating the problem into an ordinal three-class classification task: Low, Moderate, and High. This better conceptualized the nature of the severity of stress and led to more stable learning with improved accuracy and higher F1-scores.

The final ordinal Random Forest model had the best overall performance, which clearly shows the importance of problem definition alongside model choice. Moreover, the creation of a basic application in Streamlit served to prove the potential relevance of the trained model for the early detection of students who may be experiencing higher levels of stress.

In conclusion, the coursework has verified that supervised machine learning algorithms, with the right preprocessing and design, are indeed useful for uncovering information about student well-being in a learning environment.

4.2 Real-World Applicability of the Solution

A real-world challenge, an educational one was addressed in this report. The solution proposed offers how to tackle this challenge through a data driven approach for identifying students who may be experiencing stress in their life. There are many inconsistencies in the detection of stress in students through traditional methods such as self-reporting or teacher’s observation. They can be delayed and also subjective.

The proposed system has many practical relevance in the real educational environments which is possible due to analysis of multiple contributing factors. The potential real-world relevancies are listed as follows:

- The early identification of students who are at risk of high levels of stress before any serious academic or mental health consequences occur in their life.
- The conscious support for counsellors and teachers to help them make accurate and informed decision relating to students.
- The reduction of depending on single-factor detection methods that are inconsistent and not viable in actual life.
- The immense improvement of student’s wellbeing through supportive intervention that can be achieved in time.
- The contribution to better performance and engagement in regard to academics, ultimately enhancing the overall learning experience of students.

Since a survey-based data will be used to implement the solution, it makes these solutions much more practical and effective, as it does not require intrusive methods of data collection.

4.3 Further Work

In any project, there are always potentials for further works that can be implemented. While this coursework does the job of presenting a very relevant solution, there are still some future developments and enhancements within this project that can be made. These works are listed below:

- In the future, there can be implementation of the proposed system with the help of real machine learning models to help evaluate performance empirically.
- There can be experiments with additional classification algorithms or ensemble techniques to help increase the accuracy of predictions being made.
- There could be use of cross-validation and class imbalance handling techniques to make the model strong.
- In the near future, there could be incorporation of longitudinal data to analyse trends in stress.
- The integration of explainable AI technique such as SHAP, could definitely be used to improve the transparency and trust in using the system.
- A web-based application could also be developed in the future that allows real-time assessment for several education institutions, making it scalable.

All these extensions could very well be applied in the future in order to make this system much more valuable, functional and usable.

Bibliography

IBM, n.d. *How supervised learning works.* [Online] Available at: https://www.ibm.com/think/topics/supervised-learning
[Accessed 14 12 2025].

IBM, n.d. *What is classification in machine learning?.* [Online] Available at: https://www.ibm.com/think/topics/classification-machine-learning#684929709
[Accessed 14 12 2025].

Paiva, U. et al., 2025. Prevalence of mental disorder symptoms among university students: An umbrella review. *Neuroscience & Biobehavioral Reviews.*

Malebari, A. M. et al., 2024. Prevalence of depression and anxiety among university students in Jeddah, Saudi Arabia: exploring sociodemographic and associated factors. *Frontiers in Public Health,* Volume 12.

Papadogiannis, I., Wallace, M. & Karountzou, G., 2024. Educational Data Mining: A Foundational Overview. *MDPI.*

Li, K. C., Wong, B. T.-M. & Liu, M., 2024. A survey on predicting at-risk students through learning analytics. *International Journal of Innovation and Learning.*

Pataca, A. O. et al., 2025. Use of machine learning for predicting stress episodes based on wearable sensor data: A systematic review. *Computers in Biology and Medicine,* Volume 198.

Ovi, M. S. I., Hossain, J., Rahi, M. R. A. & Akter, F., 2025. Protecting Student Mental Health with a Context-Aware Machine Learning Framework for Stress Monitoring. *arxis.*

Yeler, K. & Sürücü, S., 2025. Classification of Student Stress Levels Using Machine Learning Methods. *IJANSER,* Volume 9.

Tariq, R. et al., 2025. Explainable artificial intelligence for predictive modeling of student stress in higher education. *National Library of Medicine.*

Kanevsky, J., n.d. *Research Gate.* [Online] Available at: https://www.researchgate.net/figure/Graphic-representation-of-supervised-machine-learning-In-supervised-learning-original_fig1_301688300
[Accessed 15 12 2025].

21kschools, 2025. *Top 7 Causes of Stress in Students And How to Manage Them.* [Online]
Available at: https://www.21kschool.com/in/blog/causes-of-stress-in-students/
[Accessed 15 12 2025].

Pallathadka, H., n.d. *Machine learning and educational data mining.* [Online] Available at: https://www.researchgate.net/figure/Machine-learning-and-Educational-Data-Mining_fig1_353611098
[Accessed 15 12 2025].