6DATA007C – Final Project Report

Clustering and Predicting IIT Student Stress Levels Using Machine Learning

Student: Arafath Riyas (20211613)

Supervisor: Mr. Austen

This report is submitted in partial fulfillment of the requirements for the

BSc (Hons) Data Science and Analytics

at the University of Westminster

School of Computer Science & Engineering

University of Westminster

Date: 03/07/25

Declaration

This report has been prepared based on my own work. Where other published and unpublished source materials have been used, these have been acknowledged in references.

Word Count: 10,613

Student Name: Arafath Riyas

Date of Submission: 03/07/25

Use of Generative AI

In this assessment, I used Generative AI, QuillBot, and Colab Gemini to help with brainstorming ideas, improving my writing, and fixing bugs in the code.

Abstract

Student stress is a growing concern in higher education, affecting both academic performance and overall well-being. This project aimed to identify patterns of stress and predict the stress levels of undergraduate students at the Informatics Institute of Technology (IIT) based on data-driven methods. By integrating clustering and predictive modelling, the study attempted to provide a comprehensive understanding of student stress and generate actionable insights that could be applied by the institutions

The data was collected using a structured survey focusing on the key stress-related factors, such as academics, finances, social life, coping strategies, and environmental factors. Once the data was cleaned and preprocessed (dealt with missing values, normalized, and renamed features), exploratory analysis was performed to get an overview of the distribution of variables and their interrelationships.

The K-Means clustering algorithm was applied to categorise students based on their stress responses, resulting in three significant clusters: low stress, moderate stress, and high stress. This unsupervised approach enabled natural pattern discovery without prior labels. The logistic regression was then applied to predict the membership of the cluster by individual responses. Although the model was simple, it offered strong performance and interpretability, making it suitable for institutional use.

Visualizations were used to help interpret and communicate results. The primary stress predictors included academic stress, financial concerns, and personal coping skills. The combination of clustering and classification offered an in-depth understanding of student stress.

Although the study focused on a single institution and a medium-sized dataset, the methodology provides a scalable framework for other universities to explore similar problems in student well-being. Future steps could involve expanding the dataset, testing more complex models, and incorporating real-time or longitudinal data to monitor it continuously.

Acknowledgements

I’m truly grateful to my supervisor, Mr. Austen, for his guidance and encouragement throughout this project. A special thanks to Mr. Hassan, my co-supervisor, for always being approachable and offering thoughtful feedback that helped me stay on track. I’d also like to thank Mr. Rajitha from the SRU, whose insights into the psychological side of the topic really helped me understand the issue more deeply. Finally, I’m thankful to my friends and colleagues for their support, ideas, and motivation along the way, it made the process a lot smoother and more enjoyable.

Table of contents

List of Figures

List of Tables

1. Introduction

Chapter overview

This chapter presents the research problem and justifies the need for a machine learning-based investigation into student stress. It outlines the growing issue of stress in university students with particular reference to the Informatics Institute of Technology (IIT) in Sri Lanka, a group where little research currently exists. The chapter defines the aims, objectives, and methodology of the study and clearly presents the scope of the project, its technical and contextual limitations.

1.1 Problem statement

Student mental health has become a growing issue in higher education around the world. Among the many challenges that students face, stress is consistently identified as one of the most prevalent and influential factors affecting student's academic progress and overall well-being (Beiter et al., 2015). University students face a variety of challenges, such as demanding coursework, tight deadlines, financial problems, and expectations from family and friends. The transition from the strict setting of school to more independent university settings can increase these stress levels, especially as students are frequently expected to balance multiple responsibilities with limited support systems.

While many studies have been conducted on student stress, most of them focus on specific institutions or make general conclusions that do not consider the unique environmental, cultural, and academic environments in which the students live and study. While in reality, stress level and their root causes differ between institutions due to their differences in university culture, academic demands, and availability of support systems. Even in a single university, a student's degree program, year of study, and personal circumstances may all have a significant impact on their stress experiences. This highlights the complexity and shows that universal, one-size-fits-all methods to student mental health may be insufficient to satisfy the different needs of student groups (Reavley & Jorm, 2010).

One of the biggest challenges faced by universities is recognising at-risk students early, before stress begins to affect academic performance or evolves into more severe mental illness (Eisenberg, Golberstein, & Hunt, 2009). The majority of universities continue to use wellbeing approaches mainly based on reacting, intervening only when problems are visible through declining grades, or frequent absenteeism. Not only does this late response limit the opportunity to intervene early, but it also places burdens on already overwhelmed mental health services. Consequently, numerous students who need help are left without adequate support.

While many universities around the world struggle with identifying and addressing student stress, the challenge feels even more pronounced in places where little local research exists. The Informatics Institute of Technology (IIT) in Sri Lanka is a clear example. Although IIT recognizes how important student mental health has become, there's still a real lack of detailed, data-driven understanding about what stress actually looks like for its students. To date, no studies have comprehensively studied the unique pressures IIT undergraduates face, whether those pressures come from their specific programs, their year of study, or the workload they're managing. Without such insights, IIT, like many other institutions, is often limited to broad, one-size-fits-all approaches that may overlook individual needs.

This study intends to close the gap by applying machine learning approaches to identify various stress clusters among IIT students and determining the key factors that contribute to each cluster. By considering these patterns, it becomes easy to identify which groups of students are more likely to experience higher levels of stress based on their individual characteristics. The insights obtained can help IIT build more proactive, targeted wellbeing attempts tailored to the specific requirements of each group, moving away from broad and reactive approaches to student care.

In summary, the absence of focused research on IIT students, coupled with the rising importance of mental health in education, forms the foundation of this project. Through the combination of clustering and prediction, the project aims to provide actionable insights into the stress dynamics within IIT and lay the groundwork for future improvements in student support systems.

Summary of Research Gaps in Student Stress Detection

Research Area	General Research Landscape	Gap in IIT Context
Stress in Higher Education	Widely researched in global and regional contexts	No targeted study conducted within IIT
Detection Approaches	Mostly reactive (e.g., grade drop, absenteeism)	Lack of early, data-driven risk identification
Methodology	Often uses biometric or clinical data	No ML models applied to structured survey responses
Intervention Design	One-size-fits-all models common	No segmentation or group-specific profiling
ML Use in Education Mental Health	Emerging use of clustering and prediction models	Not applied in local academic contexts like IIT

Table 1 Summary of Research Gaps

1.2 Aims and Objectives

Aims

The primary aim of this project is to design and evaluate a data-driven system that can identify and interpret stress patterns of undergraduate students at the Informatics Institute of Technology (IIT), Sri Lanka. Machine learning techniques will be employed to achieve this, particularly clustering algorithms to categorize students into distinct stress groups and classification models to identify the key factors contributing to stress within each category. These insights will be obtained from analysing responses collected through a structured student survey.

By combining two types of machine learning, unsupervised for discovery and supervised for prediction, the study achieves two main goals: it identifies stress differences among student groups and finds the most significant stressors. Lastly, the research intends to provide IIT with actionable and practical insights to identify high-risk students early on, allowing for the development of proactive and targeted mental health care.

Objectives

To achieve these aims, the project is structured around the following objectives:

1. Data Collection & Preparation

1. **Design a student-friendly survey** to gather information on demographics and key stress-related areas.
2. **Distribute the survey across IIT's undergraduate student body** to collect enough diverse responses for meaningful analysis.
3. **Clean the collected data** by fixing inconsistencies and handling any missing information.
4. **Identify and manage outliers** to make sure the data is accurate and reliable.

5. **Apply feature engineering techniques** to enhance the dataset and make it more suitable for machine learning models.
6. **Explore the data through visualizations and analysis** to uncover early patterns and better understand how students are experiencing stress.

2. Model Development & Insight Generation

1. Apply and compare multiple clustering methods (**K-Means and Hierarchical Clustering**) to group students based on similarities.
2. **Find the best number of clusters** using evaluation tools like the Elbow Method and Silhouette Score for optimal grouping.
3. Train and evaluate multiple supervised learning models (**Decision Tree and Logistic Regression**) to predict which stress cluster a student belongs to.
4. **Interpret the results from both models** to get a deeper understanding of how different stress patterns show up among students.
5. **Offer practical, evidence-based recommendations** to help IIT identify and support students who are most at risk, with the goal of encouraging proactive, rather than reactive, mental health support.

1.3 Project Scope

This project analyzes and interprets stress levels among undergraduate students at the **Informatics Institute of Technology (IIT), Sri Lanka**. The scope outlines the sections that are covered (technical and contextual) and those that are not, to ensure the work remains manageable and aligned with the academic objectives.

The primary data required for this study was collected through a structured Google form, distributed exclusively among the undergraduates of IIT. The survey covers many key aspects, including demographic details and five major stress-related dimensions: academic, financial, social, emotional/personal, and environmental areas. It also includes a section about how students cope with stress. All responses collected from

the survey are anonymous to ensure confidentiality and follows the ethical research standards.

The project uses **unsupervised learning**, specifically K-Means clustering, to group students based on how they experience stress. These clusters help reveal common patterns and variations in stress across different groups within the student group. Additionally, a **supervised learning** model using a Decision Tree classifier is developed to predict which group a student is likely to fall into, based on their survey responses. This helps to identify which stress factors are most influential.

All analysis is performed using **Python** with popular open-source libraries such as **scikit-learn, pandas, matplotlib, and seaborn**. All modelling and analysis were conducted using open-source tools; no commercial machine learning platforms were employed.

It's important to note that the project does not involve real-time stress monitoring, physiological data analysis, or clinical diagnosis. It focuses solely on IIT students, so the conclusions are not intended to be generalized to university students elsewhere. Furthermore, while the study identifies ways the university could help students, it does not propose putting those suggestions into action.

Lastly, the goal is to provide IIT with a clearer, data-driven picture of how stress impacts its students and identify areas where more tailored support could make a significant difference.

Chapter Summary

This chapter introduced the project's focus on understanding and addressing student stress at the Informatics Institute of Technology (IIT), highlighting the lack of targeted research within the institution. It outlined the problem context, its aims, and the methodological approach to it, with a focus on the use of clustering and classification models. The chapter also clarified the scope of the project, defining its boundaries and areas of focus. Collectively, these elements provide a foundation for the data-driven analysis in the following chapters

2. Background

2.1 Literature review

Stress among university students has become a growing concern due to its impact on mental health, academic performance, and social well-being. Research consistently shows that long-term stress may decrease focus, lower motivation, and raise the possibility of developing anxiety or depression. In Sri Lanka, this issue is worsened by severe academic pressure and socioeconomic expectations, particularly among students at competitive universities such as the Informatics Institute of Technology.

Traditionally, stress has been measured using psychometric tools such as the Student Stress Inventory (SSI), Depression Anxiety Stress Scales (DASS), and Perceived Stress Scale (PSS). While widely used in student research to assess emotional, academic, and physical stress, they rely on self-reported responses; the results may be influenced by response bias, which arises when students are hesitant to reveal psychological problems or may not fully recognise or express what they're feeling.

To address these issues, researchers are increasingly turning to machine learning (ML) methods, which offer more detailed and data-driven analysis. Machine learning (ML) algorithms can reveal hidden trends, group students based on shared stress indicators, and more precisely anticipate risk levels. This review explores previous research using three major approaches: clustering techniques, supervised classification models, and hybrid or ensemble methods.

1. Clustering Techniques for Stress Profiling

Unsupervised learning methods, such as clustering, have been proven valuable in identifying hidden patterns in student stress data. Dewi and Dwidsamara (2021) used the K-Modes clustering algorithm to classify students into three groups based on categorical survey responses. Each cluster represented a distinct primary stressor, such as academic overload, financial pressure, or anxiety about future careers.

In another study, Sarkar (2024) used K-Means clustering and linear regression to create a Cumulative Stress Index (CSI). The findings revealed a negative relationship between stress and academic performance, allowing for early identification of students at risk and improved chances of timely support.

Similarly, Ghildiyal (2025) used Principal Component Analysis (PCA) and clustering to develop mental health profiles that ranged from high-stress, low-coping students to those with greater emotional stability. These kinds of approaches reveal how clustering can support more tailored mental health efforts for students.

2. **Supervised Predictive Models**

Supervised classification models, such as Decision Trees, Random Forests, and Support Vector Machines (SVM), are commonly used to predict stress levels from labeled data. Patil and Chavan (2019) employed Decision Trees to identify key stress indicators such as workload and time management, achieving an accuracy rate of 83.5%.

Deena et al. (2024) examined Naïve Bayes, Logistic Regression, and SVM algorithms and discovered that Naïve Bayes achieved the highest accuracy of 90%. Their research reveals how supervised learning can provide meaningful insights that helps in early intervention for students.

Arias et al. (2024) conducted an in-depth assessment of 249 studies and discovered that SVM, K-Nearest Neighbors (KNN), and Random Forest are the most widely used classifiers in stress research, especially when combined with physiological data and validated scales such as PSS or SISCO.

On the other hand, deep learning models like Long Short-Term Memory (LSTM) networks have been applied to detect stress in real time using mobile or wearable data. For instance, Luo et al. (2024) developed Branched CALM-Net, which achieved an F1 score of 87% using smartphone data such as heart rate and sleep habits. While these models perform well, they require continuous data collection and device access, which may not be feasible for all institutions. In settings like IIT, more interpretable models, such as Decision Trees, remain a more feasible and accessible option.

3. **Hybrid and Ensemble Approaches**

Some studies have also focused on combining different models to achieve a balance between accuracy and interpretability. Doma et al. (2024) achieved 92.41% accuracy with a hybrid system that used five classifiers — Decision Tree, Random Forest, SVM, Logistic Regression, and XGBoost. For enhanced performance, the approach included using SMOTE to balance the dataset, along with feature selection and hyperparameter optimisation.

These hybrid models are especially beneficial in academic settings where transparency and personalized feedback are valued. For instance, grouping students using K-Means and then using a Decision Tree model can help identify broader stress groups while additionally highlighting individual predictors, allowing counsellors to provide more tailored support.

Relevance to the Current Study

Despite progress in stress detection using physiological sensors and deep learning models, applying such systems remains challenging, especially in resource-limited educational contexts like Sri Lanka. The majority of the studies that were examined rely on wearable technologies, mobile sensing, or large-scale real-time data collecting, which might not be feasible for universities like the Informatics Institute of Technology (IIT).

Given these limitations, this study offers a more practical and interpretable approach using structured survey data focused on academic and behavioral characteristics. This approach enables stress analysis without the use of specialist instruments or invasive data collection. The study uses K-Means clustering to uncover hidden patterns in the dataset, followed by a Decision Tree classification model to forecast stress levels in a straightforward and easy to interpret manner.

The methodology draws inspiration from previous research by Dewi and Dwidsamara, Sarkar, and Doma et al., while the techniques are tailored to the local academic context. The study's goal is to provide a scalable and context-aware system for early stress detection that can be practically implemented at IITs and other similar institutions to enhance student well-being.

Summary of Studies on Student Stress Detection

Author	Technique Category	Methods Used	Key Contribution / Insight
Dewi & Dwidsamara (2021)	Clustering	K-Modes	Segmented student stressors into 3 clusters based on categorical data.
Sarkar (2024)	Clustering + Regression	K-Means + Linear Regression	Created a stress index and showed inverse correlation with academic performance.
Ghildiyal (2025)	Clustering + Classification	PCA + K-Means + Classifiers	Grouped students by mental health levels using ML insights.
Deena et al. (2024)	Classification	Naïve Bayes, LR, DT, SVM	Naïve Bayes outperformed others with 90% accuracy.
Filippis & Foysal (2024)	Feature Analysis + Prediction	Random Forest	Ranked stressors; psychological factors like self-esteem were strongest.
Mohd & Yahya (2018)	Comparative Classifiers	Logistic Regression vs ANN	ANN showed higher accuracy (71.8%) vs LR (62.5%) for predicting depression.

Doma et al. (2024)	Hybrid Ensemble	DT, RF, SVM, XGBoost, LR	Achieved 92.41% accuracy using an ensemble model.
Luo et al. (2024)	Deep Learning	Branched CALM-Net (LSTM variant)	F1 score of 87% in mobile sensor-based stress detection.
Arias et al. (2024)	Review / Meta-analysis	249 studies analyzed	RF, SVM, KNN most frequently used; physiological data and PSS scales common.

Table 2 Summary of Studies on Student Stress Detection

2.2 Review of methods and applications

This section examines the various machine learning approaches utilized in student stress detection, focusing on the algorithms used, strengths and limitations, and how they shaped the course of this work. As research in this area continues to shift from traditional self-reported surveys toward more data-driven, ML-based approaches, the choice of methods becomes especially important, especially in environments such as the Informatics Institute of Technology (IIT), where resources may be limited. The section discusses how various machine learning algorithms have been used in earlier work and lays the groundwork for the methodological decisions made in this study.

2.2.1 Machine Learning Algorithms for Stress Analysis

- **K-Means Clustering**

K-Means is one of the commonly used unsupervised learning methods, especially for partitioning numerical data into separate clusters based on similarity. In the context of stress analysis, K-Means proves helpful in stress analysis because it identifies hidden groupings of students with similar stress profiles, especially when predetermined labels are unavailable. The algorithm is highly regarded for its speed, scalability, and simplicity. However, it assumes centroid cluster formation and is affected by

the centroid's initial placement. Several studies, notably Sarkar's (2024), employed K-Means to group students based on behavioural, academic, and lifestyle characteristics. This demonstrates its practical application in stress classification without the requirement for complex infrastructure or labelled datasets.

- **K-Modes Clustering**

When dealing with categorical data, such as that found in survey-based psychological and academic research, K-Modes is a natural extension of K-Means. Instead of using calculations meant for numbers, it measures how different categories are from each other in a way that fits non-numerical data. Dewi and Dwidsamara (2021) used K-Modes to analyse survey data and identify various groups of students, each suffering a dominant stressor. The technique is generally lightweight and simple to understand, making it suitable for non-technical stakeholders. However, it requires careful cluster selection and may struggle with high-dimensional or mixed-type data.

- **Decision Trees**

Decision Trees (DTs) are supervised learning algorithms that classify data based on a sequence of decision rules and thus are very easy to interpret and suitable for educational settings. Their straightforward branching format allows researchers and administrators to understand the reasoning behind it, which is a critical issue in student mental health environments. Patil and Chavan (2019) proved the efficiency of Decision Trees in accurately identifying key stress-related variables, such as workload and time management.

- **SVM, Naïve Bayes, and Logistic Regression**

A number of traditional supervised categorisation models are still widely used in the studies of educational stress. Naive Bayes is especially useful in low-resource environments because it is computationally simple and robust. In contrast, SVMs perform well in cases where classes are well separated and are resistant to overfitting. However, they are sensitive to kernel choice and can have trouble with noisy or large-

scale data. Logistic Regression provides a simple and interpretable model but has limited capacity to model complex and non-linear relationships.

According to Deena et al. (2024), Naive Bayes produced the most optimal results in comparison with SVM and Logistic Regression. A different study by Mohd and Yahya (2018) demonstrated that Logistic Regression achieved 62.5 percent accuracy, whereas an Artificial Neural Network (ANN) achieved 71.8 percent, indicating that more complicated models can occasionally work better when resources are available.

- **Deep Learning: Branched CALM-Net**

Real-time prediction of stress using mobile sensor data and passive behavioral tracking has been possible using deep learning methods, especially Long Short-Term Memory (LSTM) networks. Branched CALM-Net is a personalised multitask learning model that was proposed by Luo et al. (2024) and adjusts to the behaviour pattern of individual students. The system achieved an F1 score of 87% outperforming most of the conventional classifiers. Although the findings are promising, these models are highly reliant on the availability of continuous, high-quality sensor data and regular device use, conditions that are often difficult in a typical learning environment, especially in developing countries such as Sri Lanka.

2.2.2 Applications and System Implementations

Machine learning systems for student stress monitoring increasingly rely on wearable devices and mobile sensors to capture real-time behavioural and physiological data. In a review of 249 publications, Arias et al. (2024) observed that most studies employ supervised models, particularly Random Forest, SVM, and K-Nearest Neighbors (KNN), and often with standardized assessment tools, such as the Perceived Stress Scale (PSS), or SISCO Inventory, to organize their input data.

Nonetheless, most of them have been implemented in resource-rich academic settings. Not many are intended to be used in an environment where wearable sensors are not easily obtainable and the available computing facilities are limited.

This highlights the need for alternative, low-cost approaches that rely on accessible data sources like structured surveys, particularly in academic environments with limited technological infrastructure.

Comparative Table of ML Algorithms

Algorithm	Type	Strengths	Weaknesses	Example Study
K-Means	Unsupervised	Simple, scalable, interpretable	Sensitive to initialization, numeric only	Sarkar (2024)
K-Modes	Unsupervised	Supports categorical data	Needs careful K tuning	Dewi & Dwidsamara (2021)
Decision Tree	Supervised	Transparent, easy to interpret	Can be overfit without pruning	Patil & Chavan (2019)
Random Forest	Supervised	Accurate, stable, handles complexity	Less interpretable than DT	Filippis & Foysal (2024)
SVM	Supervised	High margin separation, accurate	Poor with noisy/large datasets	Deena et al. (2024)

Naïve Bayes	Supervised	Fast, efficient, good for small data	Assumes independence between features	Deena et al. (2024)
Logistic Regression	Supervised	Easy to implement and deploy	Limited by linear assumptions	Mohd & Yahya (2018)
ANN	Supervised	Learns complex patterns	High compute demand, less explainable	Mohd & Yahya (2018)
CALM-Net (LSTM)	Deep Learning	Personalized, accurate in real time	Needs sensor data, not feasible everywhere	Luo et al. (2024)

Table 3 Comparative Table of ML Algorithms

While deep learning models offer high predictive accuracy, they require significant data, infrastructure, and expertise. On the contrary, less complex algorithms, such as the K-Means and Logistic Regression, are more interpretable and easier to implement and therefore more applicable in an academic setting of limited resources. This study considers these benefits to propose a low-cost survey-based stress detection model tailored for institutions such as the Informatics Institute of Technology (IIT). By focusing on accessible data and transparent algorithms, the approach provides an effective method for early detection of student stress without the need for complex technological systems.

2.3 Review of tools

This section outlines the main programming tools, libraries and environments that were applied in the development of the student stress classification system. The selection of

tools is a key factor in the success of any machine learning pipeline, influencing stages from data preprocessing and modelling to evaluation and deployment. The tools selected in this research were determined by their compatibility with structured survey-based data and their ease of use, which are crucial factors in implementation in an institution such as the Informatics Institute of Technology (IIT).

2.3.1 Programming Language

Python was selected as the primary programming language for this project because of its simplicity and powerful ecosystem of machine learning and data analysis libraries.

- Strengths: Extensive library availability (e.g., Scikitlearn, Pandas, Matplotlib), free of charge, and simple language for quick development.

- Limitations: Slower runtime compared to compiled languages like C++ and a limited session limit.

2.3.2 ML Libraries and Frameworks

Scikit-learn

K-Means clustering and Decision Tree classification are widely used in model construction. It also supported preprocessing (e.g., StandardScaler) and evaluation (e.g., confusion matrix, accuracy score).

- **Strengths:** Simple API, reliable, well-documented.
- **Limitations:** Manual tuning is needed for some models.

Pandas and NumPy

These libraries were essential for managing the survey dataset, averaging stress domain scores, clearing missing values, and converting categorical responses.

- **Strengths:** Powerful and intuitive for structured data.
- **Limitations:** Can slow down with larger datasets.

Matplotlib and Seaborn

Used for visualizing cluster outputs, stress level distributions, and feature importances.

- **Strengths:** Customizable and easy to implement.
- **Limitations:** Requires extra effort for clean formatting.

2.3.3 Development Environment

All model development and testing were carried out using **Google Colab**, a cloud-based environment with Jupyter-like features.

- **Strengths:** No local installation required, free access to GPUs, easily shareable notebooks.
- **Limitations:** Internet dependency, limited file retention without Drive integration.

Google Colab was particularly beneficial for fast revision, making it suitable for university research projects.

2.3.4 Visualization Platform

Power BI was used to develop a visualisation dashboard that transforms ML model results into actionable information. The dashboard allows stakeholders, such as lecturers or counsellors, to investigate trends in stress distribution and identify at-risk students based on clustering or categorisation results.

- **Strengths:** Professional dashboards, interactive and user-friendly.
- **Limitations:** Requires export from python environment for visualization.

Summary Table of Tools

Tool / Library	Purpose	Strengths	Limitations
Python	Core programming	Versatile, widely supported	Slower than alternatives
Scikit-learn	ML modeling (DT, K-Means)	Simple API, good accuracy, interpretable	Not suitable for deep learning
Pandas / NumPy	Data handling & transformation	Efficient, flexible	RAM-intensive on large datasets
Matplotlib / Seaborn	Visualization & EDA	High control, academic-grade visuals	Customization complexity
Google Colab	Development platform	Cloud-based, easy to use	Needs stable internet, auto-timeout
Power BI	Dashboard visualization	Interactive and user-friendly	Requires export from Python environment

Table 4 Summary Table of Tools

The combination of Python, Scikit-learn, Google Colab, and Power Bi created an effective and well-rounded set of tools to create and display a stress classification system. The tools were chosen because they allow fast prototyping, correct ML modelling, and easy visualisation. Collectively, they offered a practical, scalable approach for this study.

3. Legal, Ethical, Sustainability and other considerations

This section describes the ethical, legal, and social responsibilities of creating a machine learning-based stress detection system. It addresses the most important topics of data privacy, informed consent, algorithmic transparency, and environmental impact. The research focuses on the rights of the participants, the reduction of harm and compliance with the institutional and national rules, which is a key to ethical conduct when dealing with sensitive student data.

3.1 Legal Considerations

When dealing with stress-related information in the educational setting, it is essential to strictly adhere to the data protection and privacy rules. In this study, no personally identifiable information (PII) was gathered, however, ethical and legal compliance remained essential.

At the collection point, all data were completely anonymised; no names, student IDs, or identifiable attributes were collected/saved. As a result, individual participants cannot be traced from the dataset. Furthermore, the project aligns with Sri Lanka's Personal Data Protection Act No. 9 of 2022, which establishes standards for legal data processing and the right to personal privacy.

3.2 Ethical Considerations

Although there was no personally identifiable information (PII) in the dataset, the delicate subject of stress-related research required a strong ethical basis. Since students are especially vulnerable when discussing emotional or academic matters, respectful engagement is essential.

To ensure the well-being of the participants, several measures were taken during the data collection process. Participation was entirely voluntary, and there was no pressure to participate. Questions in the survey were well structured to be neutral and non-

invasive to minimise any potential discomfort. Additionally, participants were made aware that the system was only meant to be used as a research and supportive tool and not as a diagnostic or labelling tool

Figure 1

In terms of model ethics, Decision Trees were chosen for their interpretability, supporting algorithmic transparency and allowing users to understand how predictions were made. This helps reduce the risks of unclear decision-making that could lead to unintended consequences.

Fairness was also prioritized. The model was carefully tested to avoid bias and did not perform profiling. Clustering and classification were based strictly on academic, lifestyle, and environmental indicators relevant to stress, not personal traits.

Notably, all findings were presented as supportive, non-judgmental indicators. The system was developed to assist early intervention and improve well-being, not to evaluate or penalize students in any way.

3.3 Sustainability Considerations

Though the project is academic in nature and does not require a lot of computation, sustainability was considered when developing the project. The decision to choose energy-efficient algorithms such as Decision Trees and K-Means was based on minimizing resource consumption as opposed to more resource-intensive deep learning models. It was developed on platforms like Google Colab, which is run on shared cloud infrastructure and assists in reducing the amount of environmental impact.

The system also had a modular and reusable design, and therefore, it could be easily adapted to future research or could be used in other academic institutions with little redevelopment. This strategy is beneficial to the environment and the sustainability of the research activities in the long run.

3.4 Social and Professional Considerations

The system was not created to encourage surveillance or judgment but to encourage student well-being. To deal with possible social implications, the following considerations were taken:

- The tool values student trust and autonomy by safeguarding student data and being transparent in the prediction generation process.
- The model and analysis were framed within the Sri Lankan academic context, and thus, cultural sensitivity was prioritised in the interpretation of indicators related to stress.
- The focus on accessibility was achieved by using tools that are easily accessible, allowing the system to be replicated or used in low-resource educational settings.

Professionally, the project displays responsible development. Machine learning methods were carefully used, validated, and tested to be reliable. The aims, possibilities, and constraints of the system were made clear, supporting an ethical use and a commitment to transparency. The design and implementation process were centred on the needs of the students in a constructive and respectful way.

3.5 Security Considerations

No security issues were involved since the data did not entail personally identifiable information (PII). Nevertheless, the data was still properly secured. All files were uploaded to the researcher's personal Google Drive account, which was protected by strong credentials and two-factor authentication. Access was restricted solely to the researcher, and no information was disclosed to any other party in the process of the project.

In conclusion, the project was conducted with a specific emphasis on legal compliance, moral responsibility, and sustainability. The study addresses the standards of responsible academic research by employing transparent and safe machine learning methods, reducing risks of data collection, and placing student well-being as a priority. These considerations help ensure that the project will promote student mental health in a fair, inclusive, and ethically acceptable way.

4. Methodology

This chapter describes the methodology that has been followed to develop a hybrid machine learning model to determine the stress level of undergraduate students of the Informatics Institute of Technology (IIT). The approach is structured into several phases: data collection, preprocessing, exploratory data analysis (EDA), K-Means clustering, Logistic Regression, and an interpretation of the model results. The phases were designed to derive meaningful insights from the survey responses of students and help in the early detection of high-stress groups.

4.1 Data

4.1.1 Data Collection

The data used in this research were collected via a structured online questionnaire that was distributed to undergraduate students at the Informatics Institute of Technology (IIT), Sri Lanka. The survey was conducted between March 1st and June 10th, 2025, and aimed to capture a broad scope of factors related to student stress, such as demographic data, academic and personal stressors and coping mechanisms.

To ensure both the content validity and reliability of the measurements, the questionnaire was designed by selectively adapting items from various well-established and psychometrically validated instruments in the study of student stress. These include:

- Student Stress Inventory *(Mohamed Arip et al., 2020)*
- College Student Stress Scale *(Feldt and Koch, 2011)*
- Perceptions of Academic Stress Scale *(França and Dias, 2021)*
- Lakaev Academic Stress Response Scale *(Lakaev, 2009)*
- Academic Anxiety Scale *(Cassady, Pierson and Starling, 2019)*
- Undergraduate Stress Questionnaire *(Crandall, Preisler and Aussprung, 1992)*

Each of these tools has been tested in past scholarly works and is widely used in evaluating measures of stress, including academic pressure, anxiety, emotional wellbeing, and environmental factors. The selective combination of the relevant items

in these sources resulted in a questionnaire that had strong content alignment with established constructs, while remaining concise, accessible, and suitable for the Sri Lankan undergraduate population.

The survey was distributed online, and the participation was entirely voluntary. The respondents were assured absolute anonymity and informed of their right to withdraw at any time. The process did not collect any personally identifiable information (PII). A total of 170 responses were obtained, and the data were then exported, cleaned, preprocessed and analysed using machine learning

4.1.2 Dataset Structure

The dataset consists of 170 complete answers gathered with the help of a structured questionnaire aiming to measure several aspects of student stress. The data is organized into seven main sections, beginning with demographic information, followed by six domains of stress:

1. Demographic Information
2. Academic Stress
3. Social Stress
4. Financial Stress
5. Personal and Emotional Stress
6. Environmental and Institutional Stress
7. Coping Strategies

The dataset consists of both categorical (e.g., age group, gender, degree program, employment status) and ordinal variables, the latter mostly measured through a 4-point Likert scale (1 = Never, 4 = Always). This combination enables subtle and interpretable inputs that can be used in clustering and classification models.

Each stress-related item was adapted from established psychological tools to indicate the frequency or intensity of certain stressors. This structured design ensures that the dataset is sufficiently rich to use in exploratory analysis and machine learning.

The Variable Description Table (see Appendix/Table X) contains a detailed description of all variables, including their definitions and data types.

Variable Types and Format

The dataset includes both:

- **Categorical variables**, include gender, age group, academic year, degree pathway, and employment status.
- **Ordinal variables**, primarily measured through a **4-point Likert scale** (1 = Never, 4 = Always), were used to assess stress indicators and coping behaviours.

Summary of Stress and Coping Measures

A total of **27 Likert-based items** were used to assess stress across multiple domains. The responses demonstrated the following trends:

- Most prevalent stressors:

  "I often feel stressed about examinations or assessments" – Mean: 2.87

  "The pressure to maintain good grades affects my mental well-being" – Mean: 2.82

- Less prominent stressors:

  "Peer competition at university creates stress for me" – Mean: 1.95

  "I feel isolated or disconnected from my peers" – Mean: 1.95

A total of 29 Likert-based items were used to assess stress in several categories. The responses revealed the following trends:

Data Cleaning and Missing Values

During the initial deployment of the questionnaire, some fields were not marked as required, resulting in incomplete submissions. This was later corrected, and:

- All fields were made **mandatory** in subsequent responses.
- Earlier missing values were imputed using mean substitution to ensure the dataset's quality and consistency.
- The final dataset is fully cleaned, consistent, and ready for analysis.

4.2 Methods

Methodology Chosen

This research adopts the CRISP-DM (Cross Industry Standard Process for Data Mining) approach, which is a widely recognised framework used to support data analytics and machine learning projects. CRISP-DM provides a structured yet flexible six-phase approach: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment. Its iterative characteristic enables continuous refinement based on emerging insight, making it suitable for research-based projects such as student stress detection.

CRISP-DM was selected due to its:

- Adaptability to Complex Problems: The phased investigation approach addresses multi-dimensional challenges in student stress
- Stakeholder Alignment: Ensures continuous alignment between technical implementation and institutional needs at IIT
- Modeling Flexibility: Supports both unsupervised (clustering) and supervised (classification) learning paradigms
- Resource Optimization: Structured phases prevent redundant iterations while accommodating unexpected discoveries

Model Building Process

1. Data Preprocessing

In this phase, the dataset was cleaned and transformed to prepare it for analysis. Some of the main preprocessing procedures included renaming the questions (Q1-Q27) and appropriately labelling demographic fields. The missing values in numeric fields were handled using mean imputation, and uniform inactive responses were deleted to maintain the quality of the data. Additionally, normalisation was used to avoid the influence of scale differences on clustering and modelling results.

To improve interpretability, feature engineering was applied by grouping related questions under each stress domain. The mean value of the relevant questions was calculated for every student, resulting in six new features: Academic_Stress_Avg, Social_Stress_Avg, Financial_Stress_Avg, Personal_Stress_Avg, Environmental_Stress_Avg, and Coping_Strategies_Avg. This allowed for a more organised examination of stress in major life domains

2. Exploratory Data Analysis (EDA)

The exploratory data analysis (EDA) was conducted to learn the overall distribution and trends in the dataset. Descriptive statistics were computed for all variables, and trends were observed by visualisation of the data using boxplots and bar graphs based on gender, academic year, and field of study.

Correlation analysis was also performed to explore how different stress dimensions related to one another. This exploratory step gave useful insights into which stressors were most prevalent among student groups and helped to build ideas for the clustering step.

3. Machine Learning and Evaluation

The primary modeling method used in this research involved both unsupervised and supervised learning methods to gain a comprehensive understanding of student stress patterns. Initially, clustering was performed through K-Means and Hierarchical Clustering to group students based on similarities.

Having compared the performance and interpretability, K-Means was chosen as the more effective one due to its clearer separation of groups. The Elbow Method was used to determine the optimal number of clusters, which was confirmed by the Silhouette Score, and the solution was three clusters. These clusters revealed distinct stress patterns among students, allowing for a detailed profiling based on average stress level, gender, academic year and field of study.

Following clustering, the generated cluster labels were treated as target variable in a supervised learning process. Two models, Decision Tree Classifier and Logistic Regression, were trained on all 27 stress-related features. Model performance was evaluated using accuracy, precision, recall, and F1-score. Logistic Regression performed better in general and provided a higher level of interpretability.

The unsupervised clustering combined with supervised classification provided a systematic method for analysing student stress patterns and predicting stress group membership. The approach supported both exploratory knowledge and predictive modelling, offering a strong foundation to interpret stress variations among student groups.

4. Results Presentation

The findings were presented both visually and in textual form to make them understandable for all audiences. The dashboard featured basic charts such as cluster distributions and stress levels, which allowed for highlighting important patterns. Meanwhile, the report provided a more detailed analysis of each cluster, including demographic trends and other key performance indicators such as F1-score, recall, accuracy, and precision.

5. Tools

This section describes the technical tools and systems applied during the development of the stress classification system. The selection was based on the suitability to analyse data, ease of integration, and compatibility with machine learning processes.

5.1 Programming Language

Python was the primary programming language used in this project. Python was chosen because of its ease of use, readability, and large data science and machine learning libraries ecosystem. It provides powerful tools for processing structured survey data, constructing interpretable models and producing high-quality visualisations. Moreover, Python is commonly used in both academia and industry, which makes it a good alternative for reproducible research workflows.

5.2 Machine Learning & Data Analysis Libraries

Scikit-learn: The Scikit-learn library was used for implementing core machine learning tasks, including K-Means Clustering, Hierarchical Clustering, Decision Tree Classification, and Logistic Regression. Additionally, Scikit-learn offered the necessary preprocessing tools such as StandardScaler to normalize the features, and model evaluation metrics such as silhouette_score, davies_bouldin_score, and calinski_harabasz_score. These played a key role in validating clustering results and supporting supervised learning.

Pandas & NumPy: Pandas and NumPy were used together for data preparation. Pandas worked with tabular data in surveys and made cleaning and transformation efficient, whereas NumPy allowed quick numerical calculations and array manipulation, underlying much of the data processing.

Matplotlib & Seaborn: Matplotlib and Seaborn were used in combination for data visualization. Matplotlib provided flexibility to make simple plots, whereas Seaborn

provided visually enhanced charts such as heatmaps and pairplots, which were helpful in exploratory data analysis.

Yellowbrick: Yellowbrick was applied in clustering evaluation. It offered visual diagnostics such as the elbow method, silhouette plots, which assisted in determining the optimal number of clusters.

5.3 Development Environment

The entire modelling and analysis were conducted in **Google Colab**, an online tool by Google that offers free GPU access and cloud storage. It was selected for its user-friendly interface, eliminating the need for local installation and its seamless integration with Google Drive. Although sessions timeout after periods of inactivity, the auto-saving feature to Google Drive helped mitigate data loss.

5.4 Dashboard and Visualization

Besides in-notebook analysis, the final output was visualised in **Microsoft Power BI**. A dashboard was created to display the results of clustering, allow filtering by demographic variables and highlight stress patterns in a visual and interactive way. Although Python has strong visualisation capabilities, Power BI was chosen to make the results easy to understand by non-technical stakeholders like staff or counsellors at the university, thanks to its intuitive interface.

In summary, the use of data analysis libraries, machine learning tools, and visualization platforms offered a complete environment to perform this study. From preprocessing of data to model evaluation and stakeholder-friendly reporting, each tool played a particular role in assisting the overall objective of this study.

6. Model development

This section presents an outline of how the stress detection model was developed, covering each phase from data preprocessing and exploratory analysis to clustering, supervised classification, and model evaluation. The project was developed in Python using **Google Colab**, with the data stored and accessed via **Google Drive** for easy workflow integration.

6.1 Data Import and Preparation

The dataset used in this study was gathered through a survey and stored in the form of a CSV file on Google Drive. The code below was used to read and prepare the data in the Colab environment:

```
# Mount Google Drive
from google.colab import drive
drive.mount('/content/drive')

# Import and store in the df variable
df=pd.read_csv('/content/drive/MyDrive/Colab Notebooks/FYP/IIT_Student Stress_Data.csv')
```

Figure 2

6.2 Data Preprocessing

Preprocessing the data was essential for preparing the survey responses for clustering and predictive modelling. This phase consisted of several steps, such as column renaming, handling missing or inactive responses, addressing skewed distributions, and normalising the feature space to have consistent model performance.

Removal of non-contributory columns

The dataset included a timestamp column automatically generated by Google Forms, which recorded the time each respondent submitted their answers. This column was not relevant to the analysis of stress prediction and was therefore removed from the dataset.

```
# Drop the first column (timestamp)
df = df.drop(df.columns[0], axis=1)
```

Figure 3

Renaming Columns

The original column names, which were initially named using the text of survey questions, were renamed to allow easier manipulation of data and provide consistency. Demographic questions got shorter and more intuitive column names, whereas stress-related questions were labelled sequentially, Q1 to Q27.

The renaming was accomplished by directly assigning the new column names to the dataset

```
# Renaming the demographic columns
df.columns.values[1] = "Age"
df.columns.values[2] = "Gender"
df.columns.values[3] = "Field_of_Study"
df.columns.values[4] = "Employment_Status"
df.columns.values[5] = "Academic_Year"

# Renaming the stress columns to Q1, Q2, Q3, ...
for i in range(6, len(df.columns)):
    df.columns.values[i] = f"Q{i - 5}"
```

Figure 4

This organised renaming was informative and reduced confusion during later phases, particularly in feature engineering and visualisation.

Handling Missing Values and Inactive Responses

After renaming the columns, the dataset was examined for missing entries. No rows contained excessive missing values, and mean imputation was applied to fill minor gaps in numerical features.

```
# Taking only the columns containing the survey responses (Q1-Q27)
question_cols = [f'Q{i}' for i in range(1, 28)]

# Fill missing values with rounded column mean
df[question_cols] = df[question_cols].fillna(df[question_cols].mean().round())
df[question_cols] = df[question_cols].astype(int)

# Ensure all values stay within 1 to 4
df[question_cols] = df[question_cols].clip(lower=1, upper=4)
```

Figure 5

Besides the processing of missing values, uniform or inactive responses, where the respondents chose the same value for all questions, were removed. These entries were non-informative because they did not offer much variance. The filter was used by keeping only rows with more than one unique response across the selected questions. A total of **15 responses** were removed.

```
# Remove rows with uniform responses
initial_rows = df.shape[0]
df = df[df[question_cols].nunique(axis=1) > 1]
removed_rows = initial_rows - df.shape[0]
print(f"{removed_rows} rows removed due to uniform responses.")
```

```
15 rows removed due to uniform responses.
```

Figure 6

Skewness and Transformation

Although Likert scale data is ordinal and tends to be less prone to extreme skew, the distributions of each stress feature were still examined to identify any potential outliers. Given that the 1-4 response scale was limited, no variables had significant skewness. Consequently, the data was not transformed.

Feature Scaling

The dataset was normalised using StandardScaler (sklearn.preprocessing) before applying clustering and classification algorithms. This was necessary to remove the influence of scale difference between features and to ensure that a distance-based model like K-Means performed optimally.

```
#Scale numerical features
scaler = StandardScaler()
scaled_data = scaler.fit_transform(df[numeric_cols])
```

Figure 7

The standardisation process transformed the data to zero mean and unit variance, preventing features with greater numeric ranges from dominating the clustering process.

Correlation Analysis

A correlation matrix was created to evaluate the connection between the 27 stress-related features. This analysis ensured that there were no redundancies that would bias future clustering or classification models. With a threshold of 0.8, none of the features were identified to exceed this level of correlation. As a result, all features were retained.

```
# Calculate the correlation matrix
corr_matrix = df[question_cols].corr()

correlation_threshold_pairwise = 0.8

# list of features that would be removed
features_removed = set()

# Start from the upper triangle to avoid duplicate pairs and self-correlation
for i in range(len(corr_matrix.columns)):
    for j in range(i):
        if abs(corr_matrix.iloc[i, j]) > correlation_threshold_pairwise:
            colname_i = corr_matrix.columns[i]
            colname_j = corr_matrix.columns[j]

            features_removed.add(colname_i)


# list of features that would remain
all_features = set(question_cols)
features_kept = list(all_features - features_removed)

# Print the results
print(f"Features removed (threshold > {correlation_threshold_pairwise}): {list(features_removed)}")
print(f"\nFeatures remained: {features_kept}")
```

Figure 8

Feature Variance Analysis

In continuation of the previous step in which inactive responses (i.e., rows with identical answers to almost all questions) were dropped, this stage aimed to assess the informativeness of each individual question by calculating its variance among all respondents. The objective was to drop survey questions that did not significantly differ among students.

The variance of each of the 27 stress-related questions was calculated, and a threshold of 0.5 was used to indicate low-variance features:

```
# Calculate variance for each question
variances = df[question_cols].var()

threshold = 0.5

# Plot
plt.figure(figsize=(16, 6))
variances.plot(kind='bar', color='steelblue')
plt.axhline(y=threshold, color='gray', linestyle='--', linewidth=1.2, label=f'Threshold = {threshold}')
plt.title("Variance of Each Question (Before Feature Selection)")
plt.ylabel("Variance")
plt.xticks(rotation=45)
plt.grid(axis='y', linestyle='--', alpha=0.6)
plt.legend()
plt.tight_layout()
plt.show()
```

Figure 9

The feature variances were visualised by a bar chart with a horizontal threshold line. Although some questions displayed lower variance, none were dropped.

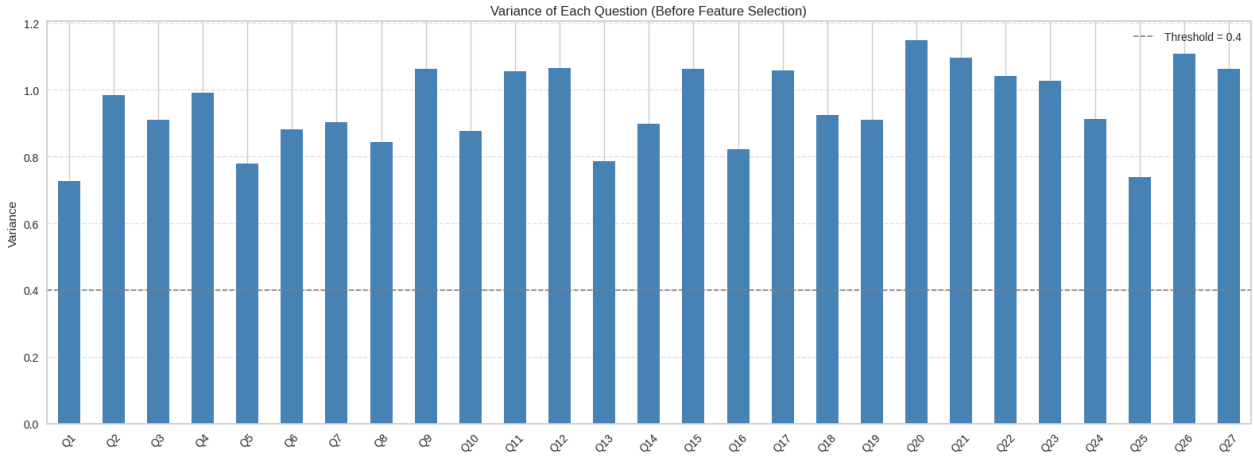


Figure 10

6.3 Unsupervised Machine Learning Model Development

This section describes the application of unsupervised learning methods to find hidden patterns in student stress data with no prior labels. Two clustering techniques, **K-Means** and **Hierarchical Clustering**, were examined to group students according to their stress responses. These models were compared and assessed to determine which provided better performance and interpretability of this dataset.

6.3.1 K-Means Clustering

The K-Means algorithm was chosen as the primary clustering algorithm in this research to cluster students based on their stress responses. It was selected due to its ease, speed and efficiency in handling larger datasets. To develop the K-Means model, the following steps were observed.

Determining the Optimal Number of Clusters

One of the most important steps in K-Means is determining the number of clusters (k). To achieve this, two techniques were used: the **Elbow Method** and the **Silhouette Score.**

The Elbow Method involves plotting Within-Cluster Sum of Squares (WCSS) against a series of potential cluster numbers. As the number of clusters increases, WCSS is likely to reduce. The most appropriate number of clusters is usually determined at the elbow point, where the rate of decline becomes steep. In the current project, the elbow was noted at **k = 3** as shown in the plot below.

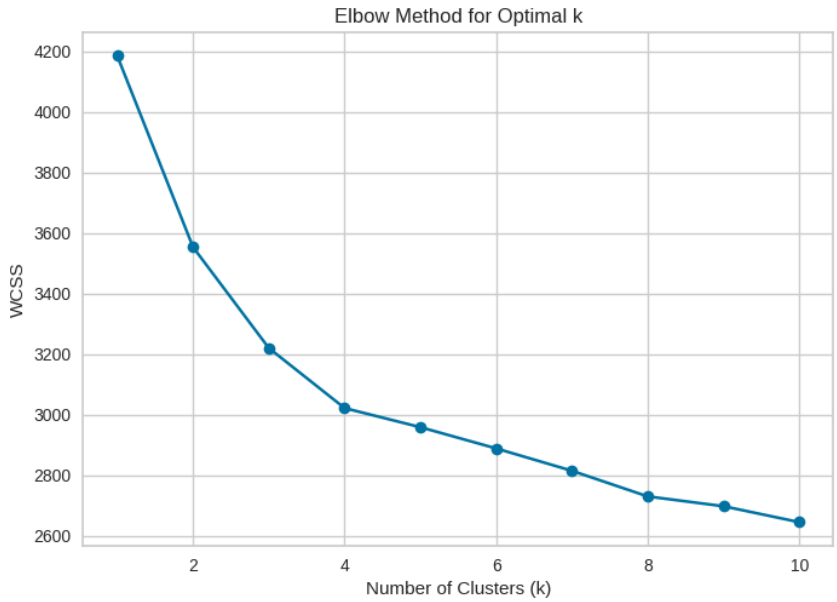


Figure 11

Simultaneously, the Silhouette Score was also computed at various k. A higher score (near 1) represents well-separated clusters, whereas a score near 0 represents overlapping groupings. The analysis revealed that k = 2 had a higher average silhouette score compared to the others.

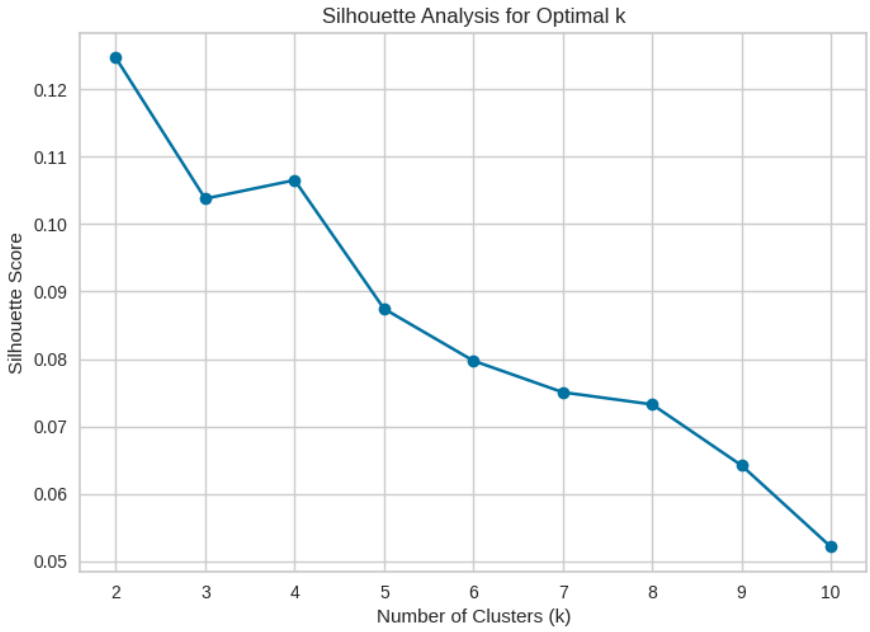


Figure 12

Evaluating 2 vs 3 Clusters

While the Elbow Method revealed that 3 clusters would be ideal, the Silhouette Score revealed a more specific structure with 2 clusters. However, due to the intent of the project to identify more specific trends in student stress, a binary classification was too simple and would probably conceal significant differences between groups.

To make a more informed decision, visual inspection and clustering measures was used to assess both cluster solutions, k = 2 and k = 3. To visualize the cluster distribution, Principal Component Analysis (PCA) was also applied. Along with the Silhouette Score, two more indicators, namely, **Davies-Bouldin Index (DBI) and Calinski-Harabasz Index (CHI)** were examined to determine the compactness of clusters.

```
# Performing K-means for 2, 3 clusters
for k in range(2, 4):
    kmeans = KMeans(n_clusters=k, random_state=42)
    labels = kmeans.fit_predict(scaled_data)

    # Scores
    silhouette = silhouette_score(scaled_data, labels)
    dbi = davies_bouldin_score(scaled_data, labels)
    chi = calinski_harabasz_score(scaled_data, labels)

    # Append results
    results.append({
        "k": k,
        "Silhouette Score": round(silhouette, 4),
        "Davies-Bouldin Index": round(dbi, 4),
        "Calinski-Harabasz Index": round(chi, 2)
    })
```

Figure 13

k	Silhouette Score	Davies Bouldin Index	Calinski Harabasz Index
2	0.1247	2.3059	27.2
3	0.1038	2.347	22.77

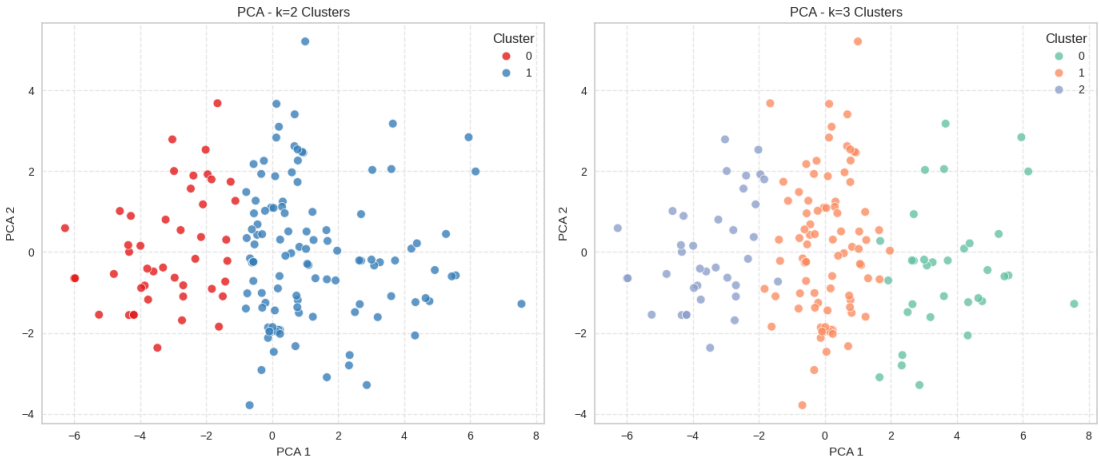


Figure 14

While the metrics were slightly in favour of k = 2, the 3-cluster solution provided more interpretable groupings. Considering these factors, the final clustering model was chosen as k = 3.

K-Means Model Development with Three Clusters

With k = 3 confirmed, the K-Means was used on the complete list of 27 stress-related features. All students were grouped into clusters according to the similarity of their responses to stress. The labels that were obtained as a result of clustering, were appended to the initial dataset as a column.

```
# Final clustering with k=3
optimal_k = 3
kmeans = KMeans(n_clusters=optimal_k, random_state=0)
df['Cluster'] = kmeans.fit_predict(scaled_data)

# Cluster means
print("\nMean values for each cluster:")
print(df.groupby('Cluster')[question_cols].mean())

# Demographic distribution
for col in ['Gender', 'Academic_Year', 'Field_of_Study']:
    print(f"\n{col} by Cluster:")
    print(pd.crosstab(df['Cluster'], df[col]))
```

Figure 15

Radar Chart for Cluster Profiling

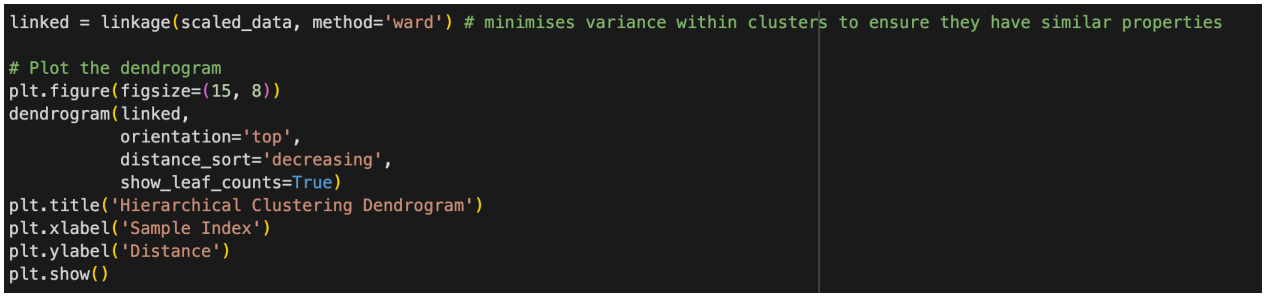
A radar chart was applied to visualize the average stress levels on academic, social, financial, personal, environmental, and coping dimensions of each cluster. This assisted in structuring the cluster profiling in subsequent analysis.

6.3.2 Hierarchical Clustering.

Hierarchical Clustering was an alternative unsupervised method that was considered to group students based on similarities in their stress responses. In contrast to K-Means, which needs a fixed number of clusters to be determined in advance,

Model Development

The Hierarchical Agglomerative Clustering (HAC) approach was applied using the Ward linkage technique, which reduces variance within clusters, producing more meaningful and compact groupings. A dendrogram was generated to visualise how the data points were combined at each step.



```
linked = linkage(scaled_data, method='ward') # minimises variance within clusters to ensure they have similar properties

# Plot the dendrogram
plt.figure(figsize=(15, 8))
dendrogram(linked,
           orientation='top',
           distance_sort='decreasing',
           show_leaf_counts=True)
plt.title('Hierarchical Clustering Dendrogram')
plt.xlabel('Sample Index')
plt.ylabel('Distance')
plt.show()
```

Figure 16

Extracting Cluster Labels

Based on the dendrogram, a 3-cluster solution was chosen to align with the K-Means setup and allow for meaningful comparison. A distance-based cut on the dendrogram was then used to extract cluster labels.

```
# Choose 3 clusters to evaluate
from scipy.cluster.hierarchy import fcluster
num_clusters_hac = 3

# Get the cluster labels by cutting the dendrogram
hac_labels = fcluster(linked, num_clusters_hac, criterion='maxclust')
```

Figure 17

Model Evaluation

The resulting HAC cluster labels were assessed using the same clustering metrics as before: Silhouette Score, Davies-Bouldin Index (DBI), and Calinski-Harabasz Index (CHI), to determine the compactness and separation of the clusters.

```
print(f"--- Evaluation Metrics for Hierarchy k=3 ---")

# Silhouette Score
silhouette_avg = silhouette_score(scaled_data, hac_labels)
print(f"\nSilhouette Score: {silhouette_avg:.4f}")

# Davies-Bouldin Index
dbi_score = davies_bouldin_score(scaled_data, hac_labels)
print(f"Davies-Bouldin Index: {dbi_score:.4f}")

# Calinski-Harabasz Index
chi_score = calinski_harabasz_score(scaled_data, hac_labels)
print(f"Calinski-Harabasz Index: {chi_score:.2f}")
```

```
--- Evaluation Metrics for Hierarchy k=3 ---

Silhouette Score: 0.0918
Davies-Bouldin Index: 2.1411
Calinski-Harabasz Index: 20.80
```

Figure 18

6.4 Supervised Machine Learning Model Development

Following the clustering step, supervised learning models were developed to predict a student's cluster membership based on their stress-related features. For this, two classification models were explored: **Decision Tree Classifier** and **Logistic Regression.**

6.4.1 Decision Tree Classifier

The Decision Tree Classifier was selected due to simplicity, interpretability, and the capacity to deal with numerical and categorical data.

Defining Target and Splitting the Dataset

The cluster labels derived by the K-Means algorithm were used to define the target variable. All 27 features related to stress served as the inputs. The data were then divided into the training and testing sets to assess the predictability of the model.

```
# Define features and target
X = df[question_cols]
y = df['Cluster']

# Train-test split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25, random_state=42)
```

Figure 19

Model Development

A Decision Tree Classifier was trained using the training dataset. For the initial version of the model, default parameters were applied to observe how the model would perform without any changes.

```
dt = DecisionTreeClassifier(random_state=42)
dt.fit(X_train, y_train)
```

Figure 20

Model Evaluation

The test set was then used to evaluate the model using standard classification metrics such as accuracy, precision, recall and F1-score. A confusion matrix was also generated to observe the misclassification

```
# Make predictions on the test data
y_pred = dt.predict(X_test)

print("--- Default Decision Tree Performance ---")
print("Accuracy:", accuracy_score(y_test, y_pred))
print("\nClassification Report:")
print(classification_report(y_test, y_pred))
```

```
--- Default Decision Tree Performance ---
Accuracy: 0.6923076923076923
```

Figure 21

Feature Importance

To find out which variables were most important in predicting the student clusters, a feature importance plot was generated.

```
# Get feature importances from the Decision Tree model
dt_feature_importances = dt.feature_importances_

feature_names = X_train.columns
dt_importance_series = pd.Series(dt_feature_importances, index=feature_names)

# Plot
plt.figure(figsize=(12, 6))
dt_importance_series.sort_values(ascending=False).plot(kind='bar')
plt.title("Decision Tree Feature Importance")
plt.xlabel("Features")
plt.ylabel("Importance Score")
plt.xticks(rotation=45, ha='right')
plt.tight_layout()
plt.show()
```

Figure 22

6.4.2 Logistic Regression

Logistic Regression is a well-known classification algorithm known for its simplicity, speed, and efficiency when dealing with binary and multi-class classification tasks. It was chosen to compare its performance against the Decision Tree Classifier.

Model Initialisation and Training

The model was initialized with a maximum iteration limit of 200 to ensure convergence and trained using the same training set.

```
# Initialize and train the Logistic Regression model
logreg_model = LogisticRegression(max_iter=200, random_state=42)

# Train the model on the training data
logreg_model.fit(X_train, y_train)
```

Figure 23

Model Evaluation

After training, the model was then used to predict the labels on the test dataset and the performance of the model on unseen data was measured using important classification metrics such as accuracy, precision, recall, and F1-score.0

```
# Make predictions on the test data
y_pred_logreg = logreg_model.predict(X_test)

print("--- Logistic Regression Performance ---")
print("Accuracy:", accuracy_score(y_test, y_pred_logreg))
print("\nClassification Report:")
print(classification_report(y_test, y_pred_logreg))
```

```
--- Logistic Regression Performance ---
Accuracy: 0.8717948717948718
```

Figure 24

A confusion matrix was generated to better understand how well the model categorised students into suitable stress groups.

```
conf_matrix = confusion_matrix(y_test, y_pred)
sns.heatmap(conf_matrix, annot=True, fmt='d', cmap='Blues')
plt.title("Confusion Matrix")
plt.xlabel("Predicted Cluster")
plt.ylabel("Actual Cluster")
plt.show()
```

Figure 25

7. Results analysis and discussion

In this section, the results of the machine learning model are discussed and interpreted. It covers a comparison of the clustering models, profiling of each student group in detail, and evaluation of the predictive models. Visualisations such as radar charts and contribution plots are generated to better understand the patterns in student stress.

7.1 Cluster Profiling & Insights

Following the implementation of the K-Means clustering algorithm using three clusters, each group was analysed according to its average stress levels on the 27 survey questions. This profiling was aimed at determining how stress levels varied among different student groups and to identify the key contributors to their stress.

Cluster 0: Low Stress Group

Cluster 0 represents the students with consistently lower stress levels in all 27 survey items, with average scores mostly within 1.3-2.1. These students experience minimal stress across key categories such as academic stress (Q1–Q5) and environmental/institutional stress (Q20–Q24), suggesting that they are either coping well or currently less affected by common stressors.

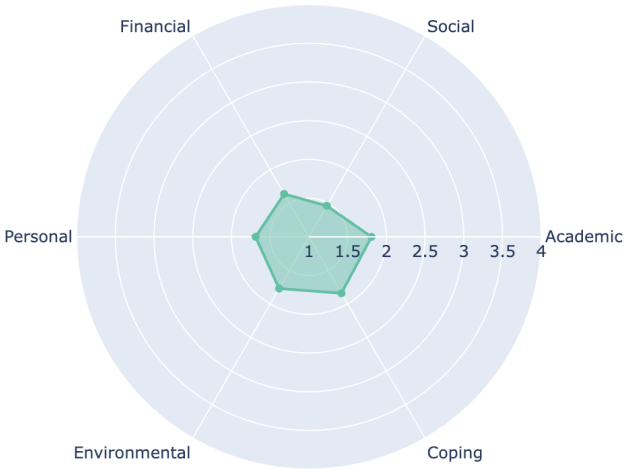


Figure 26

The cluster is predominantly male (24 out of 34), and it also includes a significant number of third-year students (17 out of 34). The third-year students likely experience lower stress levels due to reduced academic involvement, as they are typically on industrial placement. Additionally, there are 14 first-year students, a demographic that might not be completely exposed to the stress that builds up as they progress through their academic journey.

In regard to the field of study, the majority of the students belong to Computer Science and Business Information Systems, the rest are distributed across other programs. The radar and bar plots show no major spikes in any stress dimension, which further proves the low-stress profile.

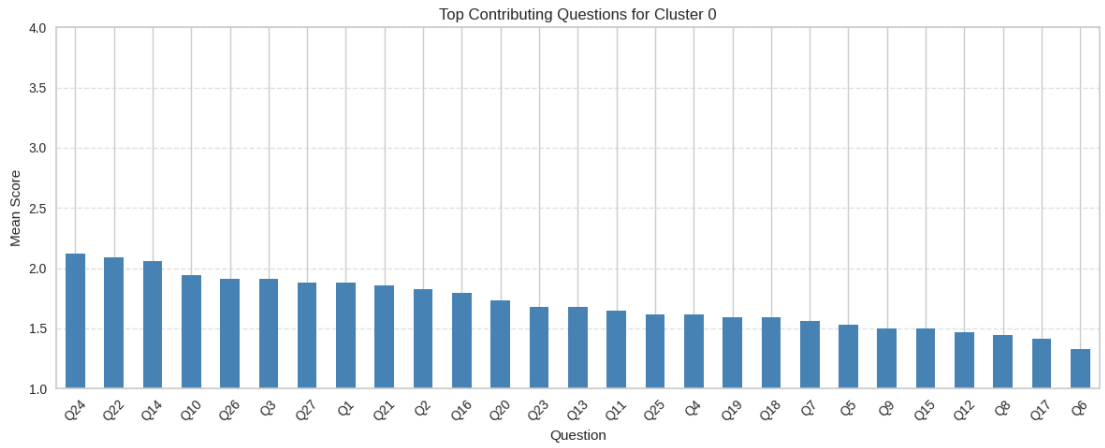


Figure 27

This group could be interpreted as a well-adjusted or low-pressure group, likely because of their academic performance or external activities during the placement year.

Cluster 1: High-Stress Group

Cluster 1 represents the group with the highest stress among the three, with consistently high mean values across all 27 stress-related questions, especially in the areas of academic (Q1–Q5), environmental (Q20–Q24), and personal stress (Q15–Q19) categories. Most questions had average scores above 3.0, suggesting frequent experiences of stress across these dimensions.

There are 46 students in this group (31 male, 15 female). The students are evenly distributed across all academic years, though slightly more concentrated in the 3rd year. This suggests that their stress may stem from non-academic sources, such as work-related pressure or transition challenges.

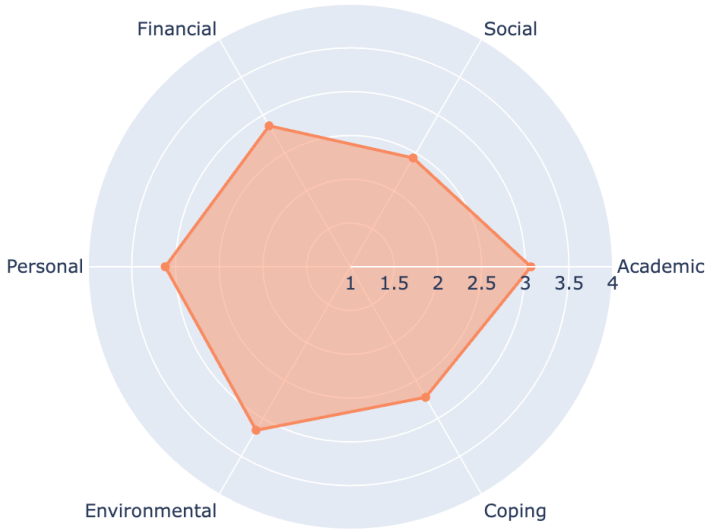


Figure 28

In regard to field of study, a diverse academic background could be observed with a notable presence of Business Information Systems (11), Software Engineering (9), and Business Data Analytics (13).

Radar and bar plot visualizations reveal stress spikes in specific areas, particularly in questions such as difficulties in concentrating (Q14), lack of motivation (Q22), emotional control (Q20), feelings of hopelessness (Q21), and insufficient sleep (Q16). These questions fall under the categories of personal and coping-related stress, emphasising this group's heightened vulnerability to maintaining their mental and emotional health.

This group may benefit most from targeted mental health interventions and personalized support programs focused on improving emotional regulation, motivation, and sleep habits. Addressing these specific challenges could help them better manage academic pressures and develop healthier coping mechanisms.

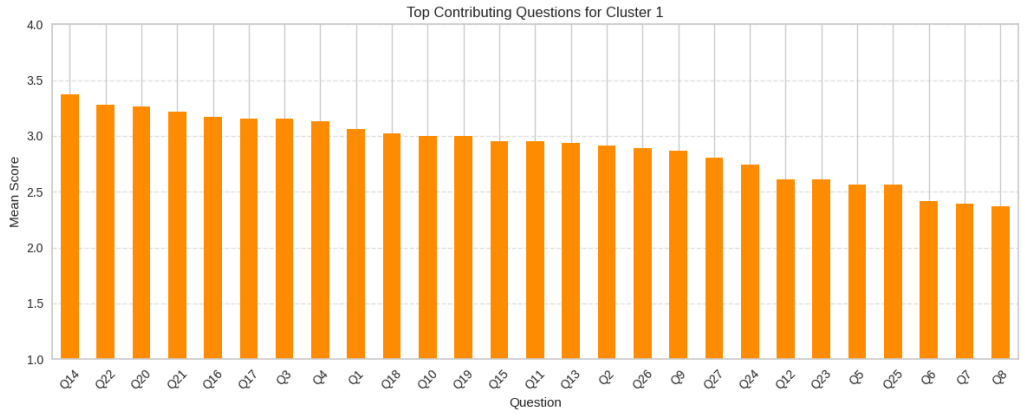


Figure 29

Cluster 2: Moderate-Stress Group

Cluster 2 represents a group under moderate stress, with average stress scores typically falling between those of Clusters 0 and 1. Notably, stress appears more visible in academic stress (Q1–Q5) and personal stress (Q20–Q24) categories, with scores ranging between 2.0 and 2.9. This indicates that while they aren't experiencing acute stress, there is a consistent presence of mild to moderate pressure.

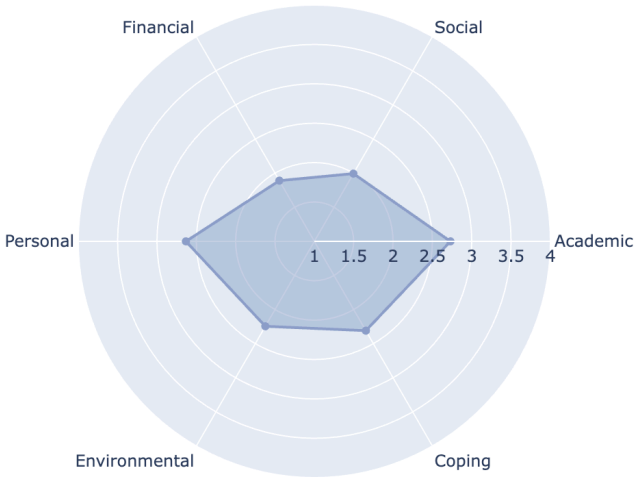


Figure 30

This cluster is the largest, with 75 students (47 male and 28 female). It contains many 3rd and 4th year students, who are likely involved in or recently finished industrial placements. Balancing both academic and work responsibilities may be a key source of stress for students in this group.

This cluster exhibits balanced representation across fields of study, with higher concentrations in computer science (18), software engineering (17), and business data analytics (17). These programs' technical and data-intensive nature probably demands prolonged cognitive effort is likely a key driver of the moderate stress levels observed.

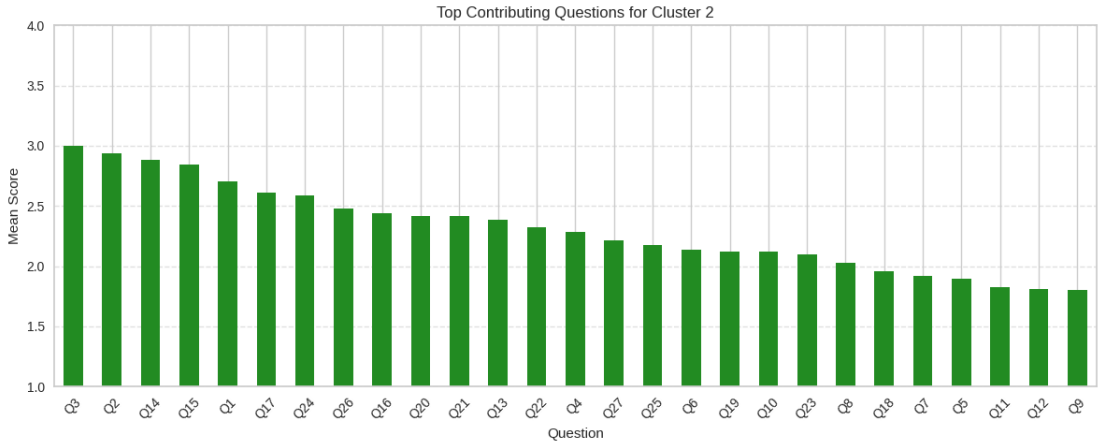


Figure 31

Radar and bar plots show more balanced stress across categories, with a few questions standing out as higher contributors: Academic Pressure (Q1), Deadline Stress (Q2), Financial Concern (Q3), Facility Access Issues (Q14), and Administrative Burden (Q15). Though this group is not as vulnerable as Cluster 1, they may still benefit from light-touch institutional support, particularly in areas of workload and time management, financial planning, and smoother access to campus services.

Stress Level	**Key Stressors**	**Majority Academic Year**	**Dominant Fields**
Low	None significant	3rd Year (on placement)	CS, BIS

Moderate	Academic & Personal	3rd and 4th Years	CS, SE, BDA
High	Academic, Enviromental, Personal	Mixed (slight 3rd year)	BDA, SE, BIS

Table 5 Cluster Profiling

Variation in Academic Stress by Academic Year

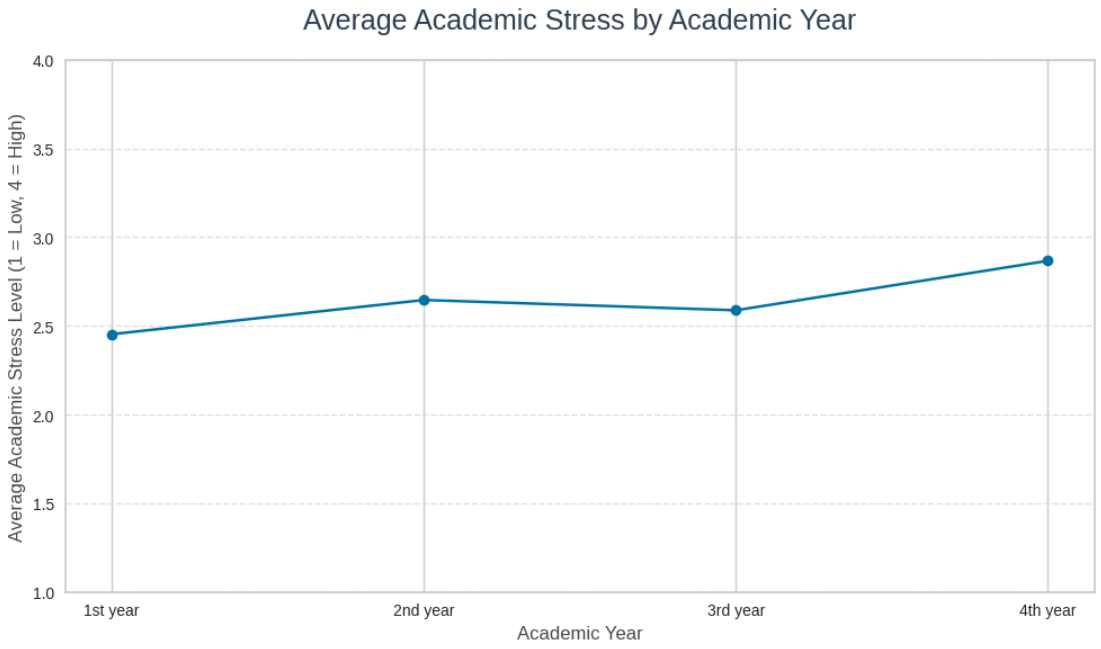


Figure 32

The Average Academic Stress by Academic Year visualisation shows a fluctuating trend in the stress level throughout the academic process. First-year students experience the least amount of academic stress, possibly because of a reduced course load and early-stage adjustment. The second year is followed by a sudden increase in stress levels, likely because of the heavier course load and higher academic demands. There is an evident decline in the third year, which aligns with the internship year, indicating that less academic activity in this year may be a factor that leads to less stress. However, stress peaks again in the fourth year, likely due to final-year projects, graduation demands, and worries about future careers.

7.2 Unsupervised Model Performance Comparison

In order to evaluate the effectiveness of the unsupervised clustering techniques, the Silhouette Score, Davies-Bouldin Index (DBI), and Calinski-Harabasz Index (CHI) were used to compare K-Means and Hierarchical Clustering.

Performance Comparison Table

Model	Silhouette Score	DBI	CHI
K-Means (k = 3)	0.1032	2.3656	22.79
Hierarchical (k = 3)	0.0918	2.1411	20.80

Table 6 Unsupervised Model Performance Comparison

In terms of Silhouette Score and Calinski-Harabasz Index, K-Means performed marginally better compared to Hierarchical Clustering; however, Hierarchical Clustering had a slightly better Davies-Bouldin Index, which indicated slightly more compact clusters. However, the differences between the two were not substantial.

K-Means was chosen as the final clustering model for additional analysis and profiling due to its interpretability, computational efficiency, and marginally superior overall clustering structure. Even so, hierarchical clustering was helpful in validating the observed cluster count and structure.

7.3 Predictive Modelling Performance

The Decision Tree Classifier (DTC) and Logistic Regression (LR), two supervised models, were evaluated to determine how well the clusters could be predicted using student responses and demographics. This was achieved using the following metrics: accuracy, precision, recall, and F1-score.

Performance Comparison Table

Metric	Decision Tree	Logistic Regression
Accuracy	0.69	0.87
Precision (Avg)	0.69	0.91
Recall (Avg)	0.68	0.85
F1-Score (Avg)	0.68	0.87

Table 6 Supervised Performance Comparison Table

Although the Decision Tree Classifier achieved an overall accuracy was 69%, its performance varied among the three clusters. Precision and recall were relatively lower in the low-stress group (Cluster 0), but it performed well for the high-stress (Cluster 1) and moderate-stress (Cluster 2) groups. This variation indicates that the model may have overfitted to dominant patterns, failing to generalise well for smaller or less distinct clusters.

In comparison, the Logistic Regression model demonstrated strong performance across all clusters, achieving a notably higher accuracy of 87%. It was particularly effective in identifying patterns within Cluster 2, with a recall of 94%, and consistently produced high precision and F1-scores. This improvement is likely due to Logistic Regression's advantages in handling high-dimensional data and modeling linear relationships more effectively than Decision Trees. Overall, Logistic Regression proved to be more suitable for this task.

7.4 Dashboard Analysis and Insights

Power BI was used to develop a simple interactive dashboard to visualize key insights from the analysis. It includes the visual representations of the cluster distributions, a radar chart, cluster cards, and the feature importance in the predictive model. The findings of this dashboard are more actionable and accessible since the stakeholders, including the university staff, can easily interpret the patterns and identify student groups at risk.

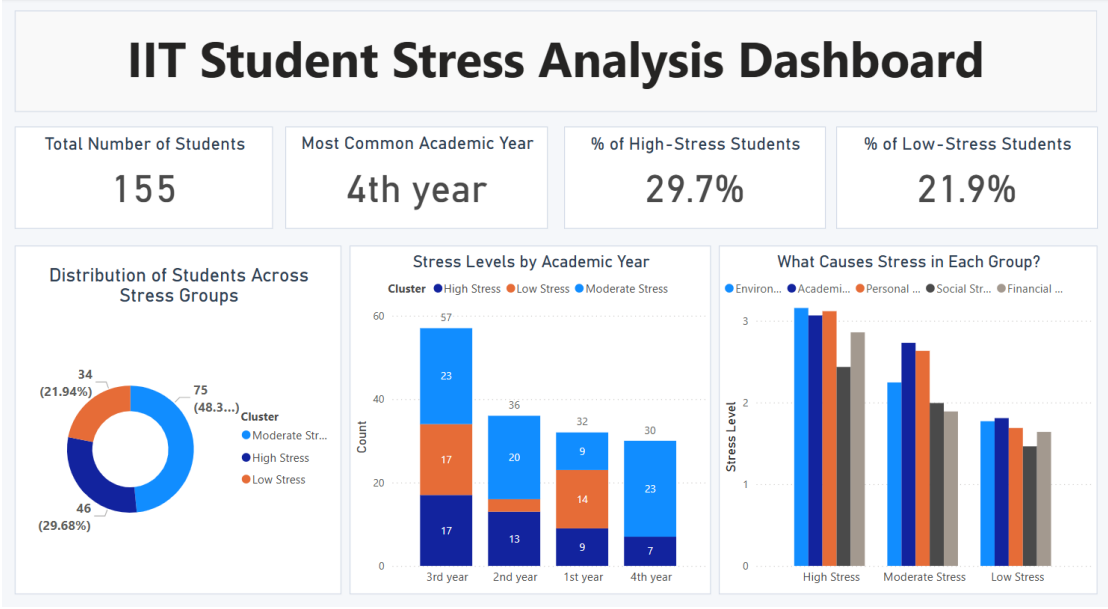


Figure 34 Dashboard

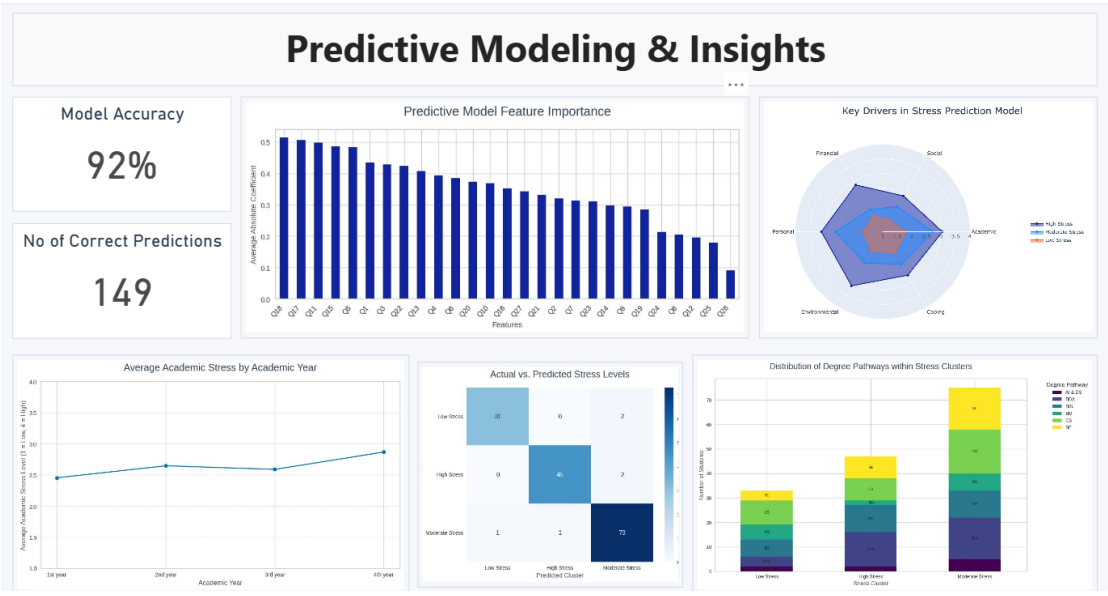


Figure 33 Dashboard

7.5 Limitations

Although this study offers insightful information about student stress, a few limitations should be acknowledged. The dataset relied on self-reported survey responses, which may be influenced by subjectivity or response bias. Furthermore, while the sample size was adequate for analysis, it may not accurately reflect the IIT student population, limiting how broadly the results can be applied within the institution.

The modelling methods employed present another limitation. Although the K-Means and Hierarchical clustering were able to provide a base knowledge on the group patterns, other clustering algorithms could reveal different structures or unknown insights. Similarly, during the supervised learning process, the research was focused on Decision Tree and Logistic Regression. More complex algorithms, such as Random Forests or ensemble models, may provide better predictive performance.

Lastly, the study lacked a live deployment system and real-time stress monitoring because it was exploratory in nature. Due to this, it has limited immediate practical application in institutional settings, however it may provide a basis for future applied research or systems.

8. Conclusions and reflections

The purpose of this project was to identify and interpret the trends of stress in undergraduate students at the Informatics Institute of Technology (IIT) using clustering and predictive modelling. The study was able to use K-Means clustering to categorise students into different stress profiles and Logistic Regression to predict the group memberships using individual responses. The results have contributed to both practical and analytical value to the academic community and student welfare efforts within the institution.

Project Summary & Findings

Based on the initial survey data, preprocessing and cleaning were performed to prepare the data for analysis. The questions were logically renamed and categorized into six dimensions: Academic, Social, Financial, Emotional, Environmental, and Coping Stress. These categories allowed for reducing dimensionality and simplifying the interpretation of results. Exploratory data analysis was used to reveal distributions and imbalances in demographic variables.

Using unsupervised learning, two and three-cluster solutions were tested and compared with each other using validation measures such as the Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index. Although the metrics were slightly in favour of the 2-cluster solution, the qualitative insights of the 3-cluster model offered a more detailed segmentation of the student stress patterns, differentiating between low, moderate, and high stress profiles.

The supervised model (Logistic Regression) further revealed which survey questions had the most impact on cluster membership. Features like academic stress, financial burden, and emotional strain were prevalent. This is especially helpful to the university staff in identifying high-risk students early.

Reflections on Methodology

The project followed a clear data science pipeline: data collection, preprocessing, exploration, modeling, evaluation, and visualization. The clustering method was useful in providing unsupervised information, and the predictive modelling supplemented this by highlighting the key variables that influenced stress. Although the algorithms were relatively simple, the interpretability and clarity of the models made them suitable for the intended use case. Moreover, the decision to use both unsupervised and supervised methods added depth to the research.

Strengths and Limitations

The balanced application of clustering and classification was one of the key strengths of the project since it provided a broader understanding of stress among students. The models were also simple to interpret and apply because of their clarity and transparency. Nevertheless, the limitations included, relatively small sample dataset, which could influence the generalizability. Furthermore, the decrease in the model performance on certain preprocessing methods, such as one-hot encoding, indicated sensitivity on feature representation.

Knowledge Gained

This project helped in developing a strong understanding of the entire data science pipeline, from data cleaning and transformation to model evaluation. It also improved the ability to balance technical choices, such as model selection and preprocessing, with practical goals such as clarity and usability. Furthermore, exploring the impact of encoding, feature selection, and clustering deepened both statistical and critical thinking skills.

Future Work

In the future, the model may be enhanced by gathering a larger and more varied dataset, which may increase its accuracy and generalizability. Experimenting with more complex models, such as ensemble methods or neural networks, may also help uncover deeper, more complex patterns in stress-related behaviour.

Furthermore, including time-based or real-timedata could allow dynamic monitoring of stress, providing a more responsive approach to monitoring student well-being. With further refinement, this project can also be used as a model by other institutions that seek to learn more about stress among their students.

Conclusion

This study demonstrated the importance of using both unsupervised and supervised methods to gain a thorough understanding of student stress patterns. By revealing hidden clusters and identifying influential factors, the study provides insight not only into the amount of stress present but also into why some students are more affected than others. The practical benefits lie in the interpretability of the models, which can guide targeted interventions and policy changes within academic settings. Overall, this project sets the framework for more data-driven approaches to student well-being and opens the way for future research using larger datasets and more complex techniques.

9. References

Beiter, R., Nash, R., McCrady, M., Rhoades, D., Linscomb, M., Clarahan, M. and Sammut, S., 2015. *The prevalence and correlates of depression, anxiety, and stress in a sample of college students*. *Journal of Affective Disorders*, 173, pp.90–96. Available at: https://doi.org/10.1016/j.jad.2014.10.054 [Accessed 15 Nov. 2024].

Eisenberg, D., Golberstein, E. and Hunt, J., 2009. *Mental health and academic success in college*. *The B.E. Journal of Economic Analysis & Policy*, 9(1). Available at: https://doi.org/10.2202/1935-1682.2191 [Accessed 25 Nov. 2024].

Reavley, N.J. and Jorm, A.F., 2010. *Prevention and early intervention to improve mental health in higher education students: A review*. *Early Intervention in Psychiatry*, 4(2), pp.132–142. Available at: https://doi.org/10.1111/j.1751-7893.2010.00167.x [Accessed 28 Nov. 2024].

Cassady, J.C., Pierson, E.E. and Starling, J.M., 2019. Predicting student depression with measures of general and academic anxieties. *Frontiers in Education*, 4, p.11. Available at: https://doi.org/10.3389/feduc.2019.00011 [Accessed 3 Jan. 2025].

Crandall, C.S., Preisler, J.J. and Aussprung, J., 1992. Measuring life event stress in the lives of college students: The Undergraduate Stress Questionnaire (USQ). *Journal of Behavioral Medicine*, 15(6), pp.627–662. Available at: https://doi.org/10.1007/BF00844860 [Accessed 4 Jan. 2025].

Feldt, R.C. and Koch, C., 2011. Reliability and construct validity of the College Student Stress Scale. *Psychological Reports*, 108(2), pp.660–666. Available at: https://www.academia.edu/127195600 [Accessed 2 Jan. 2025].

França, F.D.P. and Dias, T.L., 2021. Validity and reliability of the Perceptions of Academic Stress Scale. *Psicologia: Teoria e Prática*, 23(1), pp.1–22. Available at: https://pepsic.bvsalud.org/scielo.php?script=sci_arttext&pid=S1516-36872021000100003 [Accessed 5 Jan. 2025].

Lakaev, N., 2009. Validation of an Australian academic stress questionnaire. *Australian Journal of Guidance and Counselling*, 19(1), pp.56–70. Available at: https://doi.org/10.1375/ajgc.19.1.56 [Accessed 1 Jan. 2025].

Mohamed Arip, M.A.S., Kamaruzaman, D.N., Roslan, A. and Ahmad, A., 2020. *Student Stress Inventory (SSI): Edition 2020*. Universiti Pendidikan Sultan Idris. Available at: https://www.researchgate.net/publication/376310313_STUDENT_STRESS_INVENTORY_SSI_EDITION_2020 [Accessed 2 Jan. 2025].

Appendix I

Variable Name	Description	Type
Age	Age group of the student	Categorical
Gender	Student's gender	Categorical
Field of Study	Degree program enrolled in	Categorical
Employment Status	Employment status (part-time/full-time/unemployed)	Categorical
Academic Year	Current year of study	Categorical
Academic Workload	Feeling overwhelmed by academic workload and deadlines	Ordinal
Grade Pressure	Pressure to maintain high academic grades	Ordinal
Exam Stress	Stress related to exams or assessments	Ordinal
Class Participation Anxiety	Anxiety during class participation (discussions/presentations)	Ordinal
Relationship Maintenance	Difficulty maintaining meaningful relationships	Ordinal
Peer Competition	Stress from peer competition and comparison	Ordinal

Social Anxiety	Discomfort or anxiety in social settings	Ordinal
Isolation Stress	Feelings of being isolated or disconnected	Ordinal
Academic Financial Stress	Financial concerns affecting academic focus	Ordinal
Expense Management	Stress about managing educational expenses	Ordinal
Study Material Costs	Costs of books, materials or software causing stress	Ordinal
Family Financial Impact	Impact of family's financial condition on studies	Ordinal
Emotional Drain	Frequent mood swings or emotional exhaustion	Ordinal
Career Anxiety	Anxiety regarding career prospects or future planning	Ordinal
Sleep Disruptions	Disturbed sleep patterns due to stress	Ordinal
Time Management Stress	Difficulty managing time effectively	Ordinal
Burnout	Feeling mentally or physically drained (burnout)	Ordinal
Learning Environment Impact	Learning environment causing discomfort or stress	Ordinal

Institutional Policy Stress	Stress from institutional policies and regulations	Ordinal
Campus Mental Health Support	Availability and effectiveness of campus mental health support	Ordinal
Faculty Support	Limited or difficult access to faculty guidance	Ordinal
Administrative Process Stress	Frustration due to complex administrative procedures	Ordinal
Social Support Coping	Emotional reliance on friends or family for support	Ordinal
Physical Activity Coping	Use of physical activities to manage stress	Ordinal
Relaxation Coping	Use of relaxation techniques like meditation or breathing	Ordinal
Professional Help Coping	Tendency to avoid professional help despite high stress	Ordinal
Hobby-Based Coping	Reliance on hobbies for mental relief and stress reduction	Ordinal

Table 7 Feature Summary

Appendix II

Meeting Logbook Link:

https://docs.google.com/spreadsheets/d/1e0MnCWwE_LZZVFxMsoBXbJy2dMhZFlF2DVtDhOuh7lI/edit?usp=sharing

Questionnaire Link: https://forms.gle/Vg95RcCdVEs96EHe6

Questionnaire Responses Link:

https://docs.google.com/spreadsheets/d/1_7WN6xGlx_34ZXBhOxlHYdp-cV5yYWMdOkwuVuncj74/edit?usp=sharing

Demo Video Link:

https://drive.google.com/file/d/1-V9QTB2EFk3gdGS26cTQZMITlnq6ZqhV/view?usp=sharing

GitHub Repository Link:

https://github.com/Axafath/Arafath_w1961998_20211613/tree/main

Dashboard Link:

https://github.com/Axafath/Arafath_w1961998_20211613/blob/main/w1961998_20211613_dashboard.pbix