

**Sri Lanka Institute of Information Technology**

**Fundamentals of Data Mining - [IT3051]**

**Mini Project – 2023**

**Statement of Work Document**

Predicting Student Stress Levels

**Group – G19**

|  |  |
| --- | --- |
| IT21165702 - | Jithma V.U |
| IT21162428 - | Rathnaweera C.D |
| IT21174780 - | Dissanayaka D.M.M.I.T |
| IT21164408 - | Warunika H.P.R |

Table of Contents

[1. Background 3](#_Toc146397188)

[2. Scope of work 5](#_Toc146397189)

[3. Activities 6](#_Toc146397190)

[4. Approach 7](#_Toc146397191)

[5. Deliverables 8](#_Toc146397192)

[6. Project Plan & Timeline 9](#_Toc146397193)

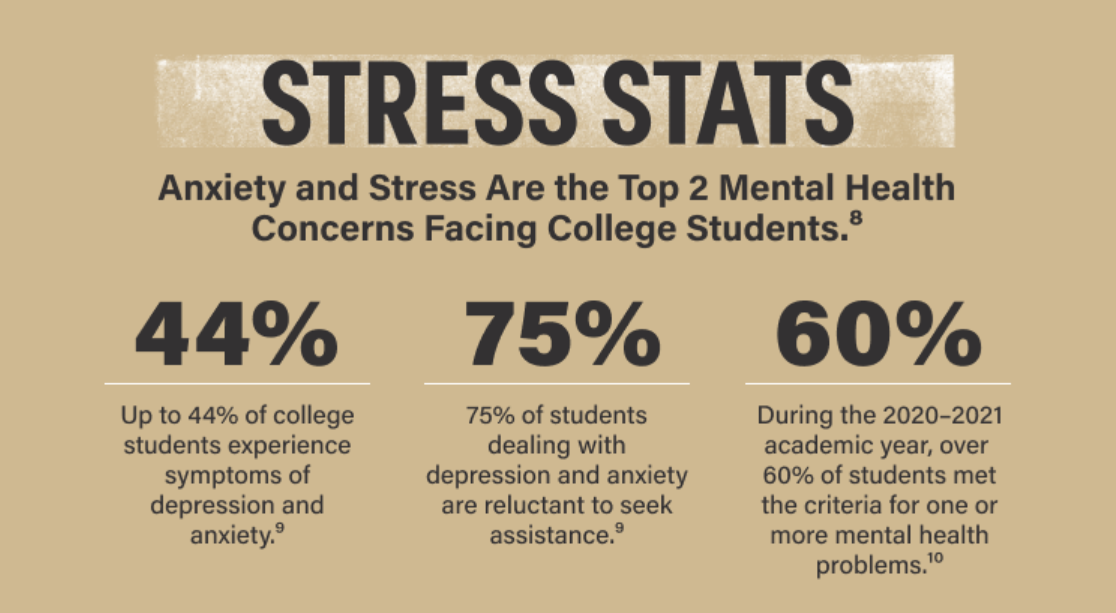
[7. Assumptions 10](#_Toc146397194)

[8. Project team, roles and responsibilities 11](#_Toc146397195)

# Background

In the modern education system, students' mental health and well-being have become increasingly critical concerns. The pressures of academic performance, social interactions, and personal challenges can lead to elevated stress levels, impacting students' overall health and academic success. To address this issue effectively, we propose a data mining project aimed at analyzing student mental health data to identify trends, early warning signs, and potential interventions. By doing so, we aim to provide timely support and interventions to students facing high-stress levels.

The background of this project revolves around the premise that data-driven insights can play a pivotal role in identifying early warning signs of stress among students. This project aims to develop a predictive model using data mining techniques to identify and anticipate student stress levels.



The motivation for this project is driven by several factors:

* **Mental Health Awareness:**

The growing awareness of the importance of mental health in academic settings and society at large has underscored the need for proactive solutions to address student stress and promote well-being.

* **Data Availability:**

The availability of extensive datasets related to student mental health, encompassing information about academic performance, demographic factors, health behaviors, and more, presents an opportunity to extract meaningful insights.

*Dataset:* [*https://www.kaggle.com/datasets/sonia22222/students-mental-health-assessments*](https://www.kaggle.com/datasets/sonia22222/students-mental-health-assessments)

* **Early Intervention:**

Timely identification of students at risk of high-stress levels can facilitate early intervention and support, potentially improving mental health outcomes and academic performance.

* **Data Mining Advancements:**

Advances in data mining and machine learning techniques allow us to harness the power of data to make accurate predictions and recommendations.

* **User-Friendly Solutions:**

Developing a software solution for stress prediction ensures accessibility to a wider range of stakeholders, including educators, counselors, and students themselves, enabling them to take informed actions.

In light of these considerations, the project's background emphasizes the potential for data mining techniques to contribute significantly to student well-being by providing insights and predictions related to stress levels. The insights gained will enable educators and support staff to intervene proactively and offer appropriate assistance to students facing elevated stress levels. This project will be a valuable tool for educational institutions and support systems to enhance student mental health support and promote academic success.

# Scope of work

This project consists of 5 main layers. Namely:

1. **User Interface Layer**

This layer is considered as the front-end of the project which allows users to interact with the system by selecting or inputting the relevant data needed for the analysis. It includes the graphical user interface (GUI) or any other facing components. This layer mainly focuses on providing a user-friendly environment for the end users to interact with the backend model of the system.

When implementing this layer, our goal is to use a simple questionnaire that contains a user-friendly interface.

1. **Data Wrangling and Data Cleansing Layer**

This layer focuses on preparing and cleaning the data for analysis. It involves tasks such as data collection, data transformation, handling missing values, and removing inconsistencies. Mainly this is the process of transforming and mapping data from the raw data form into the other format which is more appropriate and valuable for downstream analytical purposes.

1. **Data Mining Layer**

This layer is responsible for extracting valuable insights from the data. It includes various data mining and analysis techniques such as statistical analysis, machine learning, or other data-driven approaches.

The main goal of this layer is to patterns, trends, and meaningful information in the data.

1. **Model building and Analysis Layer**

In this layer, we develop and train the model. It includes tasks like model selection, feature engineering, and model evaluation. The focus is on building that can provide insights or predictions related to student’s mental health.

1. **Data Visualizing Layer**

This is the layer that is responsible for presenting the outcomes using a graphical representation in a way that is understandable to the user to get a clear view and understanding regarding the mental health of the students.

# Activities

* **Finding a real-world problem and defining a solution**

Using publicly available datasets, a real-world problem was found which is both current and relevant.

For the above real-world problem, the team was able to come up with a solution to predict the mental health of the students.

* **Data preparation, model construction, and training**

As the obtained dataset is dirty preprocessing of the dataset would occur, and the chosen data set would be cleaned (null values handled), normalized, reduced (with dimensionality reduction), and prepared to suit the implementation of the solution.

After extensive research on the problem and the solution, several models were chosen which could enable the implementation of the solution.

Then as the next step, the building of the chosen models would happen and after the models have been built, they would be trained with the prepared training set of data.

* **Evaluate the model**

As multiple models have been prepared for the solution, the most suitable one will be chosen. The evaluation of the models would happen and the best model out of the candidates would be chosen for characteristics like most accurate, least error, etc.

* **Make predictions**

Using the chosen optimal model predictions would be made to solve the business problem.

* **Front-end development**

As the final step, to release the solution to the client, a front-end application would be built. This gives a better user experience removes the technical complexity of the solution and presents the solution in a user-friendly acceptable manner.

# Approach

We will start building the model from scratch. So first we chose a dataset. Then after going through the dataset, we decided on how the dataset could be cleansed before using it to build the model. We plan to build two models using two different techniques used for multi-class classification. Then compare the accuracy of the models and proceed to build a UI to enter the properties used for the prediction and get the predicted value for the given data using the best model.

1. **Data Preprocessing**

Remove the columns without prediction power and only keep the columns that contribute to predicting students’ mental health. (Dimensionality reduction).

Remove rows with null values (null value handling).

Discretize the columns with continuous values.

Perform data normalization, reduction, and integration operations on the dataset and divide the dataset into two a training dataset and a testing dataset.

1. **Building the models**

Using the training dataset two models for multi-class classification will be built.

The following will be used to build the model

Algorithms – Decision Tree and Random Forest

Language – Python

1. **Analyzing and verifying the models**

Using the testing dataset, the models will be validated, and the best model will be chosen based on model accuracy and other metrics.

1. **Building the interface and server**

React JS will be used for the front end.

Streamlit with Python will be used for the backend.

# Deliverables

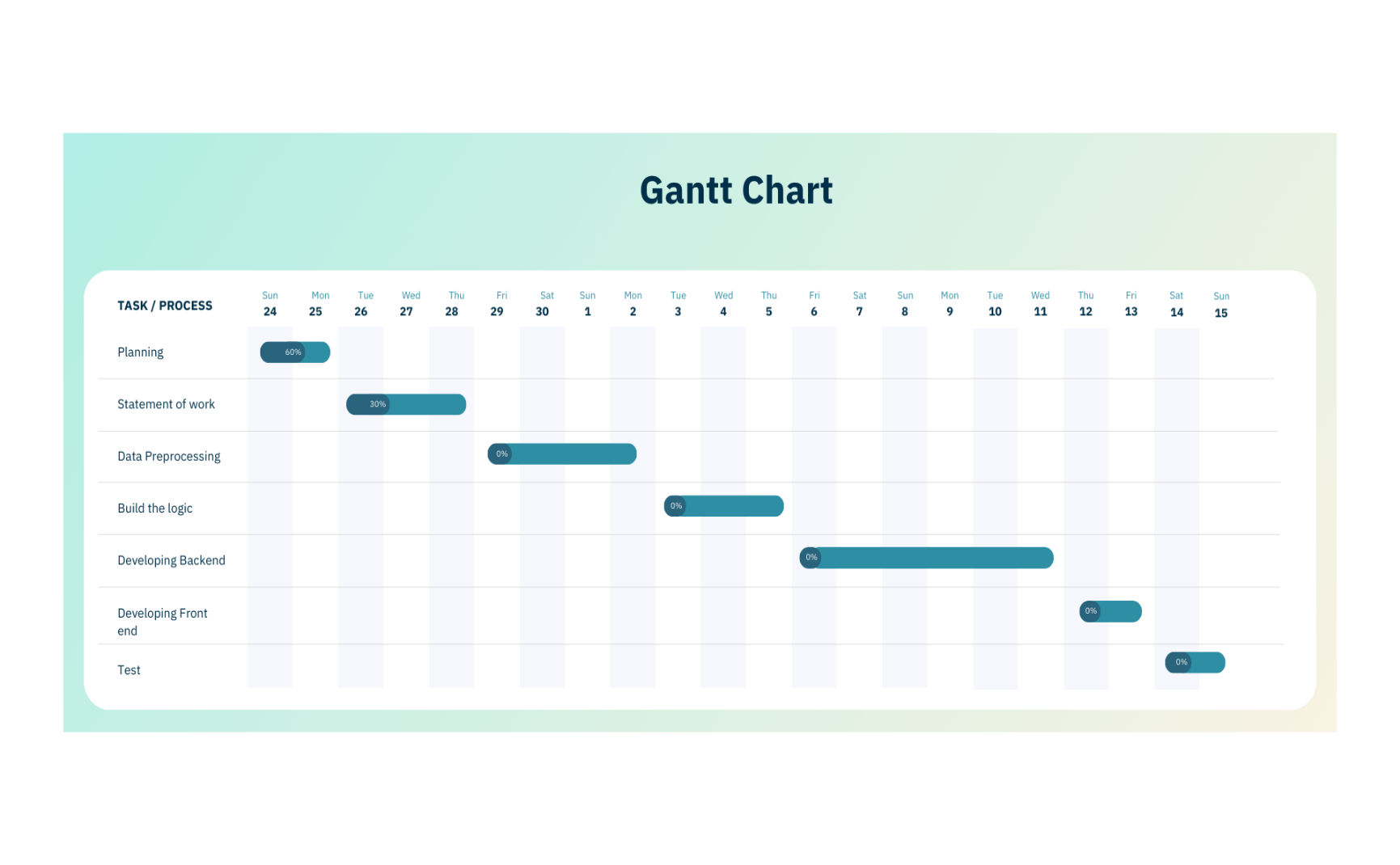
* The primary objective of this system is to gain a deep understanding of the mental health needs of students and provide effective support and resources to address those needs. By doing so, our goal is to create a system that benefits both educational institutions and the well-being of their students.
* At its core, our model and system aim to predict and identify students who may be at risk of experiencing mental health challenges. This predictive capability is vital for educational institutions to proactively address and support the mental health needs of their students, ultimately leading to improved overall student well-being and academic success.

# Project Plan & Timeline

The following project management timeline is a detailed schedule for the project. It spells out all the tasks involved and a deadline for each so that the entire team can see when individual steps will take place and when the whole project will be wrapped up.

At its core, the project timeline is an overview of the project’s deliverables laid out in chronological order. It maps out what needs to be completed before a new task can commence and keeps everything ticking along nicely.

The Gantt Chart below presents the timeline in a visual format, which means stakeholders and team members can get a quick overview immediately

****

# Assumptions

• The data is independent and identically distributed, following an unknown data distribution in which each training and test point is drawn independently.

• Initially, the entire training dataset is considered as the root.

• Records are distributed recursively based on attribute values.

• There's an assumption of no formal distributions, as it is a non-parametric model that can handle skewed and multi-modal data.

• All the source data is accurate.

• The data has not been affected by any external conditions, apart from the features contained in the dataset.

• Outliers could impact normal data, so they are removed during the data preprocessing step.

# Project team, roles and responsibilities

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Member IT**  **Number** | **Member Name** | **Member Role** | **Member Responsibilities** |
| 1 | IT21165702 - | Jithma V.U | Team Leader  Solution Developer  Solution Tester | Overall Project Coordination  Documentation  Model Implementation  UI Development  Model Testing |
| 2 | IT21162428 - | Rathnaweera C.D | Solution Developer  Solution Tester | Model Implementation  Model Testing  UI Development  Documentation |
| 3 | IT21174780 - | Dissanayaka D.M.M.I.T | Solution Developer  Solution Tester | Model Implementation  Model Testing  UI Development  Data Integration |
| 4 | IT21164408 - | Warunika H.P.R | Solution Developer  Solution Tester | Model Implementation  Model Testing  UI Development  Data Integration |