National College of Ireland

# A Deep Learning-Powered Tool for Early Detection of Student Depression with SHAP Interpretability

MSc Research Project
MSc Data Analytics

Aarthi Vijayaragavan
Student ID: X23438533

School of Computing
National College of Ireland

Supervisor: Dr. Abid Yaqoob

**National College of Ireland**
**Project Submission Sheet**
**School of Computing**

| | |
|---|---|
| **Student Name:** | Aarthi Vijayaragavan |
| **Student ID:** | X23438533 |
| **Programme:** | MSc Data Analytics |
| **Year:** | 2025–2026 |
| **Module:** | MSc Research Project |
| **Supervisor:** | Dr. Abid Yaqoob |
| **Submission Due Date:** | 28/01/2026 |
| **Project Title:** | A Deep Learning-Powered Tool for Early Detection of Student Depression with SHAP Interpretability |
| **Word Count:** | 7985 |
| **Page Count:** | 24 |

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

**ALL** internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

| | |
|---|---|
| **Signature:** | Aarthi Vijayaragavan |
| **Date:** | 28/01/2026 |

**PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:**

| | |
|---|---|
| Attach a completed copy of this sheet to each project (including multiple copies). | ☐ |
| **Attach a Moodle submission receipt of the online project submission**, to each project (including multiple copies). | ☐ |
| **You must ensure that you retain a HARD COPY of the project**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. | ☐ |

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

| **Office Use Only** | |
|---|---|
| Signature: | |
| Date: | |
| Penalty Applied (if applicable): | |

# A Deep Learning-Powered Tool for Early Detection of Student Depression with SHAP Interpretability

AARTHI VIJAYARAGAVAN
X23438533

**Abstract**

Depression among university students continues to be a major global concern, affecting not only academic performance, but also personal well-being and future potential. Conventional methods for identifying depression, such as questionnaires or clinical assessments, are often time-consuming, subjective, and not practical for early detection on a large scale. This research proposes a deep learning-based approach using a **Multilayer Perceptron (MLP)** model to predict depression from structured academic and lifestyle data, including CGPA, sleep duration, financial stress, and study satisfaction. To overcome the common "black box" issue of deep learning, the model integrates **SHAP (SHapley Additive Explanations)** to make predictions more transparent and interpretable by highlighting the factors most strongly influencing the outcome.

Experimental results show that the SHAP-enhanced MLP achieved an **accuracy of 84.6%** and an **AUC score of 0.918**, performing better than traditional machine learning models such as Logistic Regression and Random Forest. The training process required 38 seconds in a standard CPU environment, while the inference time per prediction averaged 12 milliseconds, making the model suitable for real-time applications. The global explanations via SHAP computation took 6.7 seconds and through feature importance analysis the input variables were reduced to only 10 from 77 without loss of predictive performance.

A user-friendly **Streamlit application** was developed to provide real-time prediction and visual explanations, making the model both practical and ethical for real-world use. In general, this work demonstrates how explainable deep learning can support universities and mental health professionals in identifying at-risk students early and taking meaningful preventive action.

***Keywords*— Student Depression Prediction, Deep Learning, Multilayer Perceptron (MLP), SHAP Interpretability, Real-Time AI Tool, Streamlit Interface**

# 1 Introduction

## 1.1 Background

Depression among university students is a widely reported and growing concern globally, with studies showing that the number of affected students increases each year (Hossain & Talukder 2020), (Rahman et al. 2023), (Gupta & Vishwakarma 2023). The students are under academic pressure, financial issues, lack of sleep, and even social isolation, which can take a toll on their mental health. If depression is not detected, it usually results in the decline of grades, low drive, absenteeism, and eventually a detrimental effect on one's self-esteem and career growth. The counseling services available in many colleges are not enough, so universities are now looking for digital solutions that can help them identify the students who are struggling. Machine Learning (ML) is one of the tools that can help in this matter; studies provide evidence that behavioural patterns, lifestyle habits, and academic performance can be used as indicators for predicting depression. Classic ML algorithms such as Logistic Regression, Random Forest, and AdaBoost have been very successful with survey-based datasets (Zulfiker et al. 2021), (Hossain & Talukder 2020), (Bhatt & Bhalla 2024), (Gupta & Vishwakarma 2023).

More recent studies indicate that deep learning models are even capable of achieving better performance because they can uncover intricate, non-linear correlations in mental health data. Neural networks have gained ground in the application to both numerical and survey responses(Gu 2020), (Yesudas 2022), while other researchers have gone as far as using social media posts (Redoy n.d.), (Deshpande & Rao 2017), (Arora & Arora 2019) or even EEG brain wave signals (Sarkar et al. 2022) for their experiments. However, the road to breakthroughs through these techniques is fraught with hurdles. The large number of deep learning models offers predictions without revealing the pathways that led to them, thus operating like black boxes. Some are based on data that pose serious issues in terms of scale, difficult, expensive, or even invasive to collect. Consequently, the gap between academic research and practical implementation is still very much alive. The situation has thereby brought into being a pressing need for models that are not only accurate but also transparent, feasible, and instantly acceptable by educators and mental health professionals, thus winning their trust.

## 1.2 Motivation

Numerous machine learning algorithms have shown exceptional precision in foreseeing mental health disorders, but most of them do not provide any explanation for their choices. Deep learning models are referred to as "black boxes" in several studies, they output a prediction, and the users are left guessing which factors contributed to it. This is unacceptable in an area as delicate as student mental health. The model is suspected by the counselors, psychologists, and even the students themselves to have a bias towards a particular group; hence, they would want to know the reason behind the model's belief of someone being at risk. Transparency is essential for the realisation of trust, and without it, no matter how effective a model is, it will remain in the academic realm, inaccessible to the actual students it could help. Then again, AI tech in real-life scenarios needs to score on all fronts: accuracy, explainability, and ethical responsibility.

At the same time, there is a practical problem. Certain researchers have used EEG (Electroencephalography) brain activity signals or personal social-media data (Redoy n.d.), (Deshpande & Rao 2017), (Arora & Arora 2019), which are sometimes intrusive, costly, or impractical to collect from large student populations. On the other hand, some projects yield great outcomes but never develop a tool that the students or the university can access easily. This results in a non-understanding and non-real-world impact between research and application. The study aims to build a privacy-preserving, easy-to-implement, and scalable solution by prioritising structured data such as CGPA, sleep patterns, financial stress, and study satisfaction. The intent is to develop a predictor of depression that is not only accurate but also comprehensible and trustworthy for the people.

**Research Question.**

> **How does the use of a Multilayer Perceptron (MLP) model, enhanced with SHAP-based interpretability, influence the accuracy and transparency of student depression prediction using structured academic and lifestyle data, compared to traditional machine learning models?**

1. **To what extent does SHAP improve stakeholder understanding and trust in model outputs when applied to depression prediction for educational settings?**

2. **Is it possible to use the SHAP-inflated MLP model as a real-time, easy-to-use system for providing immediate predictions of the risk of depression and explanations that are suitable for counsellors, educators and students?**

### 1.3 Research Aim and Objectives

This research intends to create an explainable deep learning model that could, with high confidence, point out those students who are at risk of depression based on their structured academic and lifestyle data. By going for an interpretability-first approach, the aim is to develop a model that is very close to the actual user and practical for usage in universities. The study is designed to follow the objectives listed below to realise its aim:

1. Create a **Multilayer Perceptron model** based on the features of students like academic performance, sleep duration, financial stress, and satisfaction with studies.
2. Apply the **SHAP** method to make it clear which factors are the reason for each prediction so that the user can understand the student's risk.
3. Assess the MLP against conventional machine learning techniques such as **Logistic Regression, Decision Tree, Naive Bayes, and Random Forest** with respect to performance metrics like **accuracy, precision, recall, F1-score, and AUC**.
4. Spot the highest-ranked indicators of depression, thus facilitating educators and counselors in recognizing the main behavioral risk patterns.
5. Develop a user-friendly and engaging **Streamlit** web application where students or counselors can type in data and see predictions along with visual indications.

This project intends to mingle high-quality predictive capability with clear-cut rationale to fill the existing gap between AI-related research and onsite mental health care. The resulting product is an early-intervention tool that carries the perks of being accurate, transparent, and therefore trusted, assisting universities to quickly identify at-risk students and subsequently engage with them confidently.

**Structure of the Document:** There are six primary sections in this report. In order to identify research gaps, the section 2 examines the literature on machine learning, deep learning, explainable AI (SHAP), and deployment systems. Section 3 describes the research methodology, including dataset selection, data preprocessing, feature engineering, system architecture, and model design. The implementation of the system, including baseline models, the MLP architecture, SHAP integration, and technical configuration, is the subject of the section 4. The results are presented and discussed in the section 5, where accuracy, AUC, confusion matrices, and training time are used to compare all models. With explanations for each prediction, the section 6 explains the Streamlit deployment and demonstrates how the model functions in an interactive real-time environment. Finally, the section 7 summarizes the main conclusions, emphasizes the significance of interpretability, talks about the limitations, and makes recommendations for future research.

## 2 Literature Review

This review critically evaluates previous research in machine learning, deep learning, explainable AI, and real-time systems for depression prediction. The review is organized into four thematic subtopics: (1) traditional ML on structured data, (2) deep learning models, (3) explainable AI using SHAP, and (4) interactive tools and system deployment. The review maps both technical progress and current gaps, all regarding the overarching research question:

**How does the use of a Multilayer Perceptron (MLP) model, enhanced with SHAP-based interpretability, influence the accuracy and transparency of student depression prediction using structured academic and lifestyle data, compared to traditional machine learning models?** .

### 2.1 Traditional Machine Learning on Structured Data

Several studies have applied traditional machine learning models to structured educational, medical, or lifestyle data to predict mental health disorders. Zulfiker et al. (2021) applied AdaBoost with feature

selection to forecast psychosocial data with 92.56% accuracy. Hossain & Talukder (2020) claimed that mental health prediction is achievable using structured student surveys using ML techniques. Baba & Bunji (n.d.) supplemented this by showing that basic classifiers like logistic regression and decision trees can fit mental health risk well.Cruz et al. (2023) applies a Naïve Bayes classifier to predict the severity of depression in university students and it was reported that the method is effective in detecting moderate and severe cases of depression with a great overall accuracy rate of 78.03%.

Other studies, such as Rahman et al. (2023) and Gupta & Vishwakarma (2023), extended the scope to encompass lifestyle data like diet, stress, and social activity in ML models for depression prediction. Alishiri et al. (2008) used logistic regression to predict mental and physical health outcomes among chronic disease patients, validating the role of structured data in medicine.

Despite these attempts, the majority of work is un-interpretable and lacks user feedback mechanisms. Such lack of model transparency is particularly concerning in sensitive contexts like student mental health, where teachers and users require not only accuracy but also trust. Such works justify the use of structured data but call for stronger and more transparent models—precisely what this work achieves by integrating MLP with SHAP.

## 2.2 Deep Learning and Mental Health Prediction

Deep learning models have been applied very frequently to depression prediction nowadays, with higher complexity and often better performance than traditional ML. Yesudas (2022) has demonstrated that neural networks perform better than classical ML for predicting depression, anxiety, and stress based on DASS-42 data. Gu (2020) implemented a deep learning-based intervention model specifically developed for mental health issues within college students.

Sarkar et al. (2022) trained Deep Learning (MLP, RNN, LSTM) vs. ML-SVM models using EEG signals to compare performance on temporal and non-linear data. Although they predominantly outperformed other models, they demanded large EEG datasets, making them unsuitable in student environments due to hardware and privacy constraints.

Bhatt & Bhalla (2024) showed that even simpler models like Random Forests and Linear Regression can also be taught patterns from survey data, though they cannot handle complex feature interactions like MLPs can. Manjunath (2022) and Bhat (2021) have used machine learning for other diseases, further proving the effectiveness of structured biomedical data in predictive tasks.

However, in these studies, deep models are often "black boxes" without explanation at the feature level. None integrate SHAP or explainable methods with neural networks on tabular education data indicating a gap that our MLP + SHAP model directly addresses.

## 2.3 Explainable AI and SHAP Integration

Explainability is critical in mental health forecasting, where black-box models hinder uptake and trust. Suhas et al. (n.d.) tackled the issue of small sample bias and rarity of events using Firth's penalized logistic regression on national mental health survey data. Though not a deep learning model, it emphasized robustness and fairness in health modeling. SHAP, which is an explainable AI technique, plays a crucial role in improving the transparency of mental health prediction systems, indicating the contribution of specific features to the model's decision making, and especially deep learning is now considered to be a reliable method in the healthcare domain due to this (Manjunath 2022).

Garg & Kaur (2025) applied SHAP to CatBoost for depression prediction, demonstrating how local and global feature attributions improve interpretability. Similarly, Chawla et al. (2002) introduced SMOTE to address class imbalance, a problem also relevant to rare mental illness diagnosis.

Although these works advance the field toward transparency, they often lack neural architectures or deployment strategies. This project builds on them by integrating SHAP with MLP to provide both performance and ethical interpretability, crucial for sensitive and imbalanced datasets like those on student depression.

## 2.4 Real-Time Interfaces and Practical Deployment

Some recent publications focus on the application of ML models via real-time or interactive interfaces. Suvarna (2020) developed a Twitter-based mental illness predictor using Streamlit but relied on shallow classifiers and text data only. Pawar & Thorat (2023) applied a basic student well-being monitor but without interpretability.A web application based on Streamlit was deployed in Nath et al. (2025) to

implement the trained model for real-time predictions and visualizations, which made the mental health screening in educational contexts practical and friendly for the end-user.

Deshpande & Rao (2017) and Arora & Arora (2019) used sentiment analysis and ML over social media data to detect depression, but did not use structured inputs or deploy real-time tools. Bhatt & Bhalla (2024) and Gupta & Vishwakarma (2023) proposed visual dashboards, improving usability but lacking explainability.

These systems reflect a growing need for ML tools that are accurate and user-friendly. However, none combined deep learning, SHAP, and live deployment for structured educational and lifestyle data. This study bridges that gap with an MLP + SHAP model implemented in Streamlit to enable real-time, interpretable depression prediction.

### 2.5 Research Niche and Contribution

Three main gaps can be seen while reviewing the literature:

1. MLPs are underutilized for the prediction of depression in structured academic and lifestyle data, as most studies rely on traditional models such as Logistic Regression, Decision Trees, Random forest, or SVMs instead of deep neural architectures ((Zulfiker et al. 2021),(Hossain & Talukder 2020),(Rahman et al. 2023),(Bhatt & Bhalla 2024),(Gupta & Vishwakarma 2023)).

2. In this study, SHAP and explainable AI techniques are less incorporated into deep learning models. While SHAP has been applied to tree-based models such as CatBoost (Garg & Kaur 2025) and some studies focus on transparent regression-based analysis (Suhas et al. n.d.), there is limited work combining SHAP with neural networks for student mental-health prediction.

3. Very few systems offer real-time, interpretable interfaces targeting student populations. Existing applications either lack interpretability (Pawar & Thorat 2023) or rely on text from social media rather than structured academic data (Suvarna 2020)(Deshpande & Rao 2017)(Arora & Arora 2019).

This study responds to these gaps by

- Applying a multilayer perceptron that has been trained on academic, lifestyle, and psychological data.
- Using SHAP to clarify both global and individual predictions.
- Deploying the model via Streamlit for instantaneous, comprehensible predictions for users.

In this way, the study offers a scalable, ethical, and interpretable diagnostic tool that is specifically designed for student mental health support as its contribution.

## 3 Research Methods

In carrying out this investigation, the CRISP-DM (Cross-Industry Standard Process for Data Mining) method was preferred. This is a generic methodology that is commonly used in making predictions and doing research based on data. The framework includes business understanding, data understanding, data preparation, modelling, evaluation, and deployment. This Figure 1 shows the overall process followed in the project, from dataset preparation to prediction and deployment.

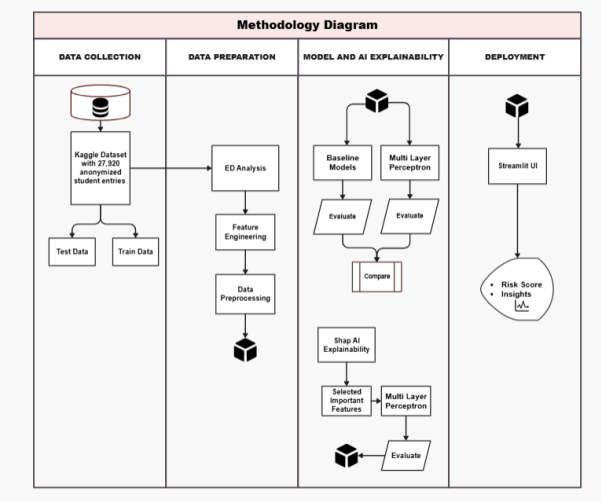


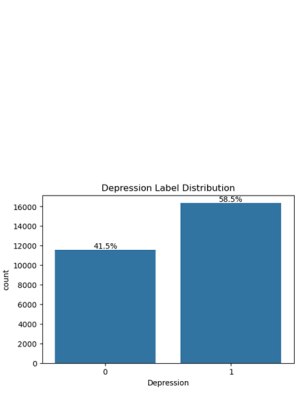

Figure 1: End-to-End Methodology of the Depression Prediction System Using Machine Learning and SHAP-based Explainability

## 3.1 Business Understanding

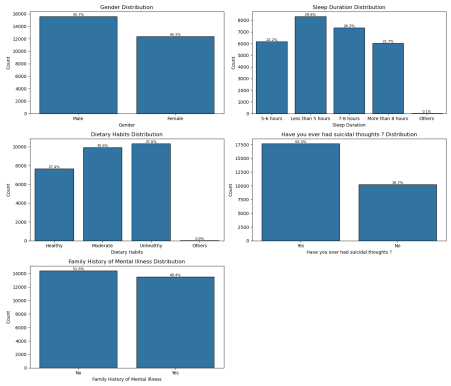
This research implements a quantitative, experimental design based on supervised machine learning. The primary objective is to create a predictive model that would be able to spot depression tendencies among students through the use of structured academic, lifestyle, and behavioural data. A data-driven approach provides the guarantee that results are reproducible, measurable, and to be evaluated objectively instead of relying on subjective judgement. The methodology, by training and testing models on actual student records, gives a systematic way to analyse mental-health patterns and assess how accurate a model is in distinguishing depression risk. This framework allows scientific validation, comparison with baseline methods, and transparent reporting of performance outcomes.

## 3.2 Data Understanding

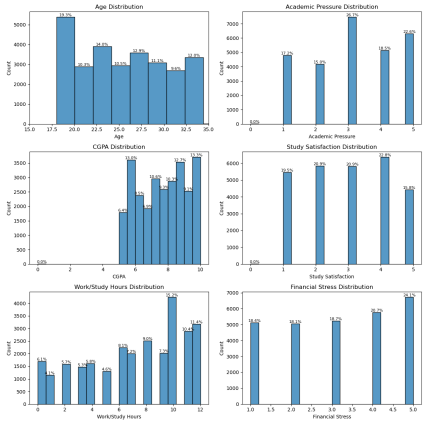
The Kaggle Student Depression Dataset was used which inclusively portrayed 27,920 anonymised student records. This dataset consisted of not only demographic variables (age, gender, city) but also of academic indicators (CGPA, study hours), lifestyle attributes (sleep duration, dietary habits), and factors affecting mental health (stress, suicidal thoughts). The dataset can be accessed for further review and download.[1]




(a) Depression Label Distribution




(b) Categorical Feature Distribution




(c) Numerical Feature Distribution

Figure 2: Overview of depression labels and feature distributions in the dataset

Initially, Exploratory Data Analysis (EDA) was performed to analyze variable distributions, spot outliers, and get a grasp of target balance (depressed vs non-depressed). Some of the methods used

[1] `https://www.kaggle.com/datasets/hopesb/student-depression-dataset`

for visualization were histograms, correlation matrices, and boxplots that all contributed to the early discovery of patterns. It was noted that class imbalance was significant, and thus methods such as resampling and threshold tuning were later applied.These Figures 2 and 3 depict the dataset composition by illustrating depression label distribution along with major categorical and numerical features in the dataset.

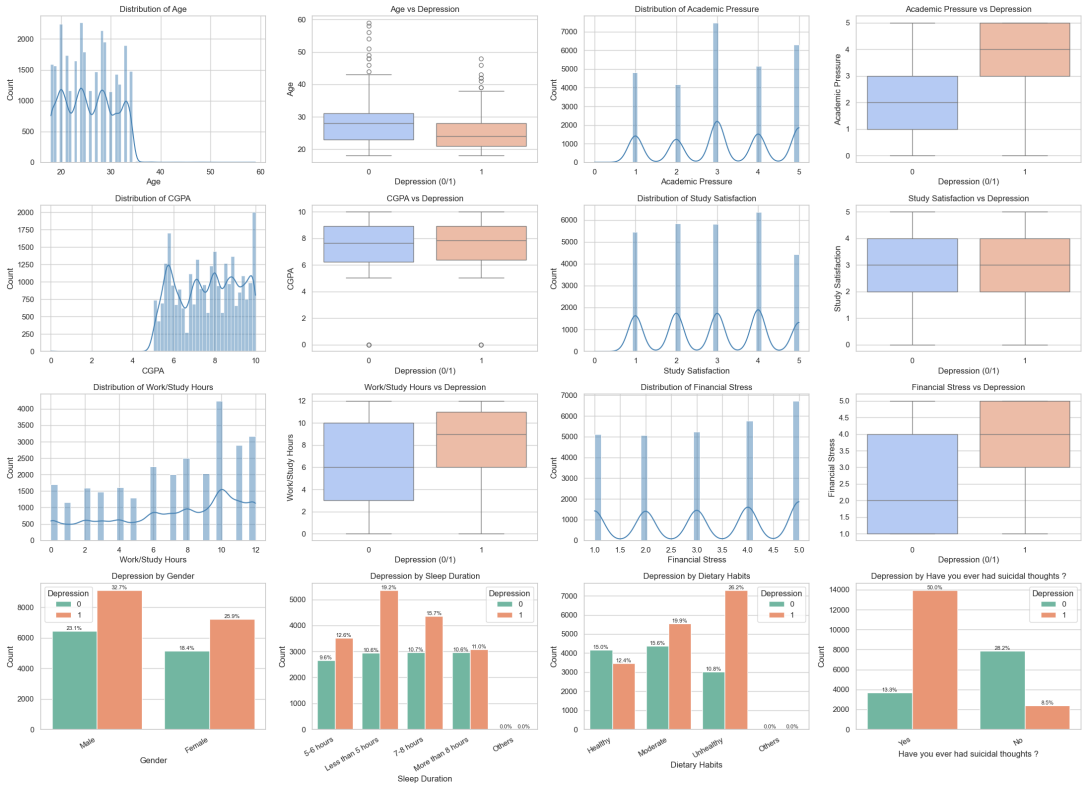



Figure 3: Distribution and Depression-Based Variation in Key Numerical Features

## 3.3 Data Preparation and Feature Engineering

A structured preprocessing pipeline was created to prepare the dataset for further processing and later modelling. Mean substitution was employed for numerical data with missing values and the mode was used for missing categorical data the methodology is that which mental-health prognosis studies based on structured survey data follow (Hossain & Talukder 2020), (Rahman et al. 2023), (Gupta & Vishwakarma 2023). For data to be compatible with machine-learning algorithms, one-Hot Encoding was applied to convert all categorical variables into numerical form, which is also the method used in similar research on predicting depression and behaviour (Zulfiker et al. 2021), (Yesudas 2022), (Bhatt & Bhalla 2024).

The numerical features were normalised through Min-Max scaling so that their value range became 0 to 1. This procedure is not only intended to stabilise the gradient-based optimisation of the MLP during training, but it is also one of the essential requirements for deep-learning models applied to datasets in psychology or biomedicine (Gu 2020), (Sarkar et al. 2022). In addition, the grouping of rare categories with very low frequency into an "Other" class was another step towards improving data quality. Noise from sparsely represented categories was thus reduced, a dilemma also described in previous studies on mental health modelling with a variety of demographic variables (Rahman et al. 2023), (Pawar & Thorat 2023).

The interaction feature of domain-driven nature, Work Pressure × Study Satisfaction was created to reflect the combined influence of workload and academic satisfaction, two factors frequently mentioned as stress contributors in students in literature (Rahman et al. 2023), (Gupta & Vishwakarma 2023). It was all done using a ColumnTransformer pipeline so that there would be consistency and reproducibility. The dataset after preprocessing consisted of 77 predictors, including the new interaction term and the transformed categorical variables.

Ultimately, the data was divided into training and testing sets through an **80:20** stratified split that kept the initial class distribution intact. This technique is in line with the recommended practices when it

comes to the fair and reliable evaluation of depression-prediction models (Zulfiker et al. 2021), (Yesudas 2022), (Bhatt & Bhalla 2024). This preprocessing workflow yielded a feature matrix that was clean and well-structured, which made it possible for the MLP model to learn optimally while preserving data integrity and interpretability.

The chosen model that forms the basis of this study was the **Multilayer Perceptron neural network**. The MLP was chosen because it could identify complicated, non-linear dependencies in academic and lifestyle data, a point supported by previous mental-health investigations(Gu 2020)(Yesudas 2022).

## 3.4 Architecture

The architecture of the suggested depression-prediction framework is unveiled in this section. The complete system comprises five primary parts: input of data, preprocessing, prediction based on MLP, interpretability using SHAP, and a Streamlit dashboard for instant interaction.

### 3.4.1 Modelling

The modelling procedure occurred in two main phases. Initially, the dataset was utilized to train a group of standard machine-learning algorithms, namely **Logistic Regression** (Baba & Bunji n.d.)(Alishiri et al. 2008), **Decision Tree** (Baba & Bunji n.d.), **Naïve Bayes**(Cruz et al. 2023), and **Random Forest**(Bhatt & Bhalla 2024). Through this approach, the performance of the deep learning technique was immediately put into a comparison frame with the machine learning models. These models were selected for their popularity in mental-health prediction studies using structured student data (Rahman et al. 2023)(Gupta & Vishwakarma 2023).

### 3.4.2 System Architecture

The overall system structure is presented in Figure 4. When data is entered either from the dataset or directly by a user, it first passes through the preprocessing stage, where missing values are filled; categorical answers are converted into numerical format, and numerical features are scaled. Once the data is cleaned, it is sent into the Multilayer Perceptron (MLP), which produces a probability score indicating the likelihood of depression.

Alongside this prediction, a SHAP explanation layer calculates how much each feature (such as sleep duration or financial stress) contributed to the result. Both the predicted risk level and the SHAP explanations are then shown in a Streamlit dashboard that users can interact with.

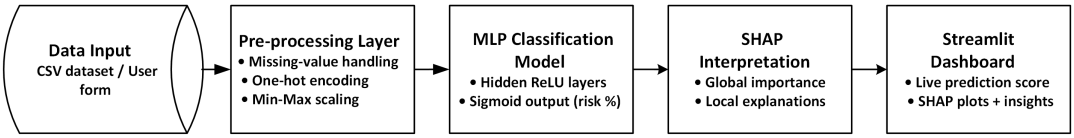



Figure 4: End-to-End Architecture of the Depression Prediction and Explainability System

The approach of having prediction and explanation in a single pipeline guarantees system accuracy and, at the same time, opening up the process to scrutiny, thus making it applicable to counselors, teachers, or students who require precise and reliable results.

### 3.4.3 MLP Architecture

The internal structure of the proposed Multilayer Perceptron (MLP) is shown in Figure 5. The network starts with an input layer that receives all encoded academic, lifestyle, and psychological features from the dataset. These inputs are then passed through three fully connected hidden layers, each using a **ReLU activation function**. ReLU is commonly used in modern neural networks because it helps the model learn non-linear relationships more quickly and avoids issues like vanishing gradients.

The first hidden layer contains **128 neurons** and is followed by **batch normalisation** and **dropout (0.4)** to stabilise learning and reduce overfitting. The second hidden layer includes **64 neurons** with **batch normalisation** and **dropout (0.3)**, while the third has **32 neurons with dropout (0.2)** for further regularisation. Finally, the output layer consists of a **single sigmoid-activated neuron**, which generates a probability between 0 and 1 indicating whether a student is likely to be at risk of depression.

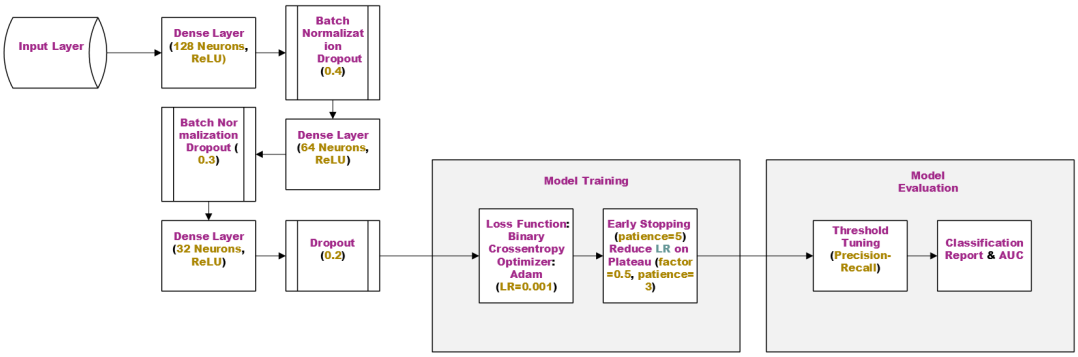



Figure 5: Architecture and Training Pipeline of the Multi-Layer Perceptron (MLP) Model

The model was trained using **the Adam optimiser** (learning rate = 0.001) and binary cross-entropy loss. Early stopping and **Reduce-LR-on-Plateau** were applied to prevent overfitting and ensure that the model could generalise to new data. Despite being a deep learning approach, this architecture is lightweight enough to be trained on a standard computer without a GPU.

The choice of an MLP is supported in previous work. Gu (2020) showed that deep learning models outperform traditional classifiers when predicting psychological outcomes in college students. Yesudas (2022) also found that neural networks were more effective than classical machine-learning models for detecting depression, anxiety, and stress. Sarkar et al. (2022) reached a similar conclusion using EEG data, where deep learning captured depression-related patterns better than SVM-based approaches. These findings support the use of an MLP to model the complex relationships found in student mental-health data.

In summary, the proposed MLP combines strong predictive ability with efficient training and good regularisation, making it suitable for practical use in academic or early-screening environments.

### 3.4.4 SHAP Interpretability Flow

Deep learning models often behave like "black boxes", making decisions without clearly showing how they arrived at the result. To solve this, the system integrates a SHAP-based explanation layer. As shown in Figure 6, the trained MLP model and its input features are passed into a SHAP explainer, which produces two types of insights:

- **Global feature importance**, showing which factors such as sleep hours, academic pressure, or financial stress have the biggest influence on predictions across all students.
- **Local explanations**, using waterfall or force plots, which explain why a single student was predicted as high-risk or low risk by revealing how each feature pushed the prediction up or down.

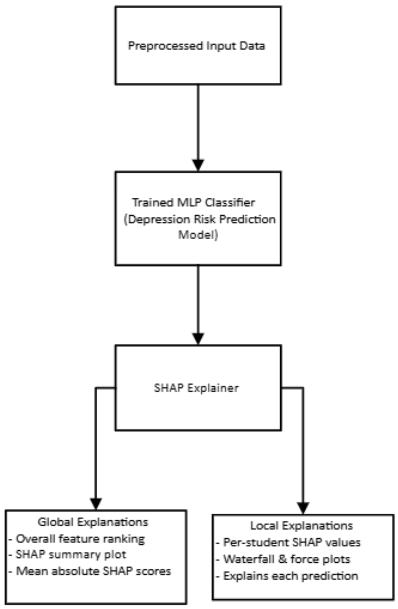



Figure 6: SHAP-Based Explainability Workflow for the MLP Classification Model

These explanations make the model's reasoning clear and easy to understand, even for users without technical knowledge. Instead of simply seeing a risk score, counsellors and educators can view the factors behind it, which supports transparent, ethical, and responsible use of AI in real mental-health settings.

## 3.5 Interpretability Using SHAP

The "Black box" is one of the common terms used to describe deep learning models as they predict without providing the steps followed to come up with the predictions. To Solved this problem, SHAP was performed on MLP's predictions. SHAP gives each feature a person's sleep duration, financial stress, or study satisfaction, for instance a contribution score that reveals how much that feature was responsible for pushing the prediction towards "depressed" or "not depressed."

Globally, SHAP reveals the most important features that are influential over the whole dataset. Locally, it justifies why a particular student was classified as high- or low-risk. The transparency provided by this mechanism allows teachers, therapists, and researchers to see the logic behind the decision made by the model, thus promoting fairness, trust, and ethical practices in the use of AI in real-world scenarios (Suhas et al. n.d.); (Garg & Kaur 2025)).

### 3.5.1 SHAP Feature Importance Results

The SHAP results provide insight into which factors contributed most strongly to the MLP's predictions. Three forms of visualisation were generated to support interpretability:

1. **SHAP Summary Plot (Beeswarm Plot)**

   This plot 7 shows both the importance and direction of influence for each feature. Key observations include:

   - "Have you ever had suicidal thoughts? = Yes" was the strongest predictor, consistently pushing predictions toward high depression risk.
   - Financial Stress, Work/Study Hours, Unhealthy Dietary Habits, and Academic Pressure showed strong positive contributions to risk.
   - Some features (e.g., Sleep Duration or Family History of Mental Illness) had moderate but consistent effects, often shifting predictions slightly toward higher risk.

- Feature colours show value magnitude (blue = low, red = high), clearly indicating how high-stress variables increased predicted depression probability.

This Figure 7 depicts the distribution of SHAP values for each feature, illustrating how high and low feature values impact the depression risk prediction.

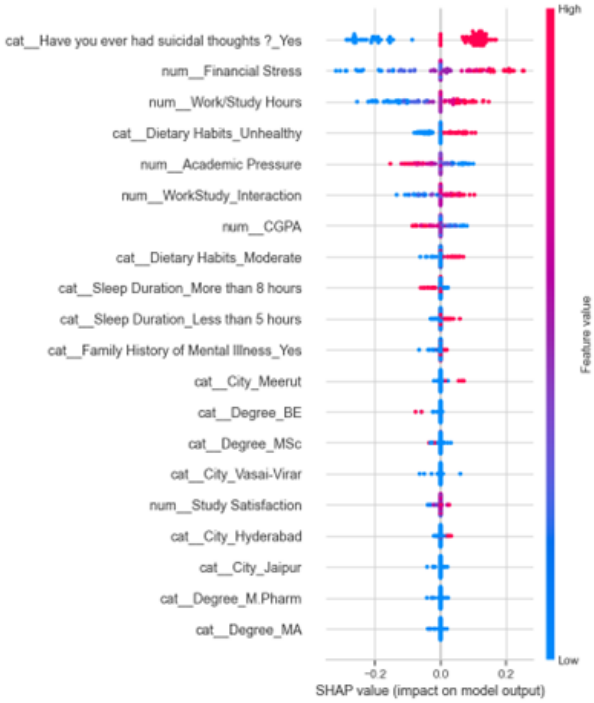



Figure 7: SHAP Summary Plot Showing Feature Influence on Model Predictions

2. **Mean SHAP Value Bar Chart**

This Figure 8 depicts the top contributing features ranked by their mean absolute SHAP values and relative percentage impact on depression prediction.

Top global predictors included:

- Suicidal thoughts (26.3%)
- Financial Stress (20.0%)
- Work/Study Hours (14.9%)
- Unhealthy Dietary Habits (6.7%)
- Academic Pressure (6.6%)
- Work–Study Interaction (6.0%)
- CGPA (3.5%)

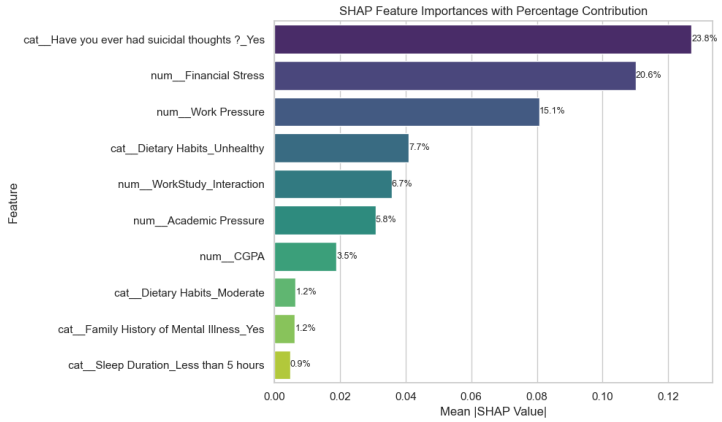



Figure 8: SHAP Feature Importance with Relative Percentage Contributions

These results align with previous findings in mental-health and behavioural-risk literature, where stress, lifestyle patterns, and high workload frequently appear as dominant predictors (Gupta & Vishwakarma 2023)(Rahman et al. 2023).

3. **Domain-Level SHAP Impact Analysis**

To interpret the SHAP results more easily, all features were classified into three domains: Mental Health, Academic, and Lifestyle. The SHAP plot at the domain level (Figure 9) depicts the contribution of each domain to the MLP's predictions.

The findings indicate quite evidently patterns, which are in line with the previous mental-health research:

- Mental Health factors proved to be the most significant ones in making predictions. Features like suicidal ideation and a family history of mental illnesses were the most contributing ones, which is in accordance with the risk indicators recognized by psychological studies (Rahman et al. 2023)(Gupta & Vishwakarma 2023).
- Academic factors turned out to be the second most significant group. The most prominent contributors were academic pressure, CGPA, and work/study hours—they were the same as those in the previous studies that marked academic stress as a key predictor of student mental issues (Hossain & Talukder 2020)(Rahman et al. 2023).
- Lifestyle factors, such as sleep and nutrition, were associated with smaller but considerable contributions. This is in line with prior research, which pointed out that lifestyle habits could be considered secondary but still relevant factors contributing to depression (Bhatt & Bhalla 2024)(Pawar & Thorat 2023).

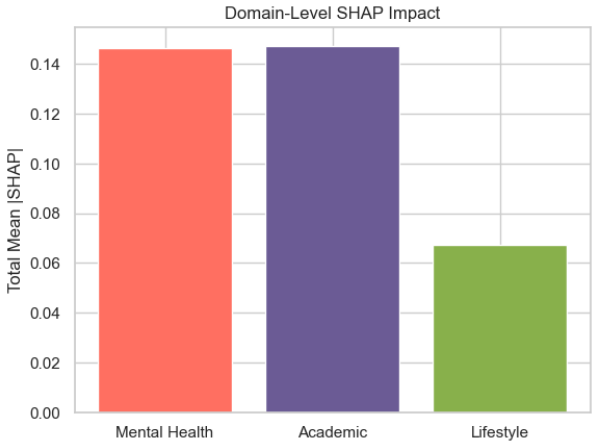



Figure 9: Domain-Level Contribution to Depression Risk Based on SHAP Values

In general, the domain-level evaluation concludes that the risk of depression is due to a mixture of mental, academic, and lifestyle stresses. The cross-domain influence not only supports the findings of the broader literature but also gives a clearer picture to the users of which category of factors is the strongest one affecting the model's output.

### 3.6 Deployment Strategy

The trained MLP model was rolled out via a **Streamlit web application**, making it user-friendly for the public, hence outside the research environment. Instead of having to use technical instruments or know the programming language, the users can easily alter the sliders and select from dropdown menus to input values like sleep duration, study satisfaction, financial stress, or coping habits. After receiving the inputs, the system immediately provides a **depression risk prediction** according to the trained neural network.

One of the primary aims of deployment was interpretability. In line with this, SHAP visualizations were integrated into the interface. The dashboard offers two kinds of explanations: global feature importance indicating which factors are the most influential overall and local SHAP waterfall plots further clarifying how individual inputs contributed to a given prediction of the student. This way, the results are made accessible to counselors, educators, and students, and not solely to data-science specialists.

The application can be run either locally or on **Streamlit Cloud**, and it neither saves nor sends personal data, hence **privacy and ethical usage are assured**. By integrating real-time prediction with simple visual explanation, the system turns into an effective tool for early detection and sensitization in schools.

## 4 Implementation

### 4.1 Setup and Data Configuration

An environment was created in **Python 3.9**, and machine-learning libraries like **TensorFlow/Keras, Scikit-learn**, and **SHAP** were helped as the means for model training and interpretation. Standard desktop computers with an **Intel Core i7** processor and 8 GB of RAM were used for all experiments. Thus, the full pipeline is run without specialized hardware, which is convenient for other researchers or educators to reproduce or extend the work.

The Kaggle Student Depression Dataset, with 27,920 anonymised student records, was the main data source consistent with prior research that uses structured academic and lifestyle data for mental-health modelling (Hossain & Talukder 2020), (Rahman et al. 2023), (Gupta & Vishwakarma 2023). The target label was set during the initial configuration, feature types (categorical and Numerical) were specified, and the dataset was split into training and testing sets in an 80:20 split, which follows common practice in similar predictive-modelling studies (Zulfiker et al. 2021), (Yesudas 2022), (Bhatt & Bhalla 2024) to allow for a balanced evaluation process. This configuration ensured that the model was trained under reproducible, well-structured conditions suitable for both deep learning and explainable AI analysis.

### 4.2 Baseline Model Implementation

To guarantee a fair benchmark for comparison, the dataset that had been preprocessed was used to train four traditional machine learning models, which were **Logistic Regression** (Baba & Bunji n.d.)(Alishiri et al. 2008), **Decision Tree** (Baba & Bunji n.d.), **Naïve Bayes**(Cruz et al. 2023), and **Random Forest**(Bhatt & Bhalla 2024) respectively. Each model underwent fitting on 80% of the training split while the evaluation was carried out on the other 20%. Where suitable, hyperparameters were tuned using grid search, such as adjusting the regularisation strength for Logistic Regression or exploring different tree depths for Decision Tree models, to ensure that each classifier performed at its best.

The performance of the models was evaluated through standard metrics which were Accuracy, Precision, Recall, F1-Score, and AUC-ROC. These metrics are not only the most utilized in mental-health prediction studies but also provide a comprehensive view of performance, especially in the case where there are significant real-world ramifications of false positives or false negatives (Rahman et al. 2023), (Gupta & Vishwakarma 2023). As an additional measure, visual tools like confusion matrices and ROC curves were produced to make it easy to see the behavior of each baseline model during the classification process.

The results of these baselines were the starting point for checking if the proposed MLP model would be able to give meaningful improvements. The use of similar classical algorithms as benchmarks in previous research on depression prediction with structured data is frequent and shows their importance in validating the merits of the more advanced deep-learning methods (Rahman et al. 2023), (Gupta & Vishwakarma 2023). Establishing these baselines ensured that any improvements achieved by the MLP or its SHAP-enhanced version were measurable, justified, and grounded in proper experimental comparison.

Table 1: Classical Models: Hyperparameters and Configuration Summary

| **Model** | **Core Hyperparameters** | **Exact Hyperparameters** | **Remarks** |
|---|---|---|---|
| Logistic Regression | L2 penalty; regularisation strength (C) tuned via grid search | max_iter=2000, penalty=l2, solver=lbfgs, C=1.0, class_weight=None | Strong linear baseline; fast, stable, and reliable |
| Decision Tree | Max depth, min samples split, and criterion (Gini/Entropy) tuned | criterion=gini, splitter=best, max_depth=None, min_samples_split=2, min_samples_leaf=1 | Highly interpretable; sensitive to overfitting |
| Naïve Bayes | Default variance assumptions; no major tuning required | GaussianNB, var_smoothing=$10^{-9}$ | Extremely fast; limited by independence assumption |
| Random Forest | Number of trees, max depth, and min samples split tuned | n_estimators=100, criterion=gini, max_depth=None, min_samples_split=2, min_samples_leaf=1, bootstrap=True | Most stable classical model; strong non-linear learning |
| Common Framework | Uniform preprocessing and scaling strategy | Encoding, feature scaling, and consistent pipeline across all models | Ensures fair benchmarking before MLP comparison |

## 4.3 MLP Implementation with SHAP Interpretability

The Multilayer Perceptron (MLP) model was used to find out if a deep learning architecture could surpass the traditional machine learning baselines in terms of predictive performance and be interpretable via SHAP. In fact, deep learning has already proved its superiority over classical methods in mental-health prediction tasks, especially when it came to representing complex behavioral and psychological patterns (Gu 2020),(Yesudas 2022),(Sarkar et al. 2022). The final network included an input layer, **two hidden layers with 64 and 32 ReLU activated neurons respectively**, and a **single sigmoid output neuron** for binary classification. To mitigate the risk of overfitting, **dropout regularization (0.3 and 0.2)** was employed, and the model was trained with **binary cross-entropy loss** using the **Adam optimizer (learning rate=0.001)**, an optimization strategy widely used in deep-learning mental health studies (Gu 2020),(Yesudas 2022).

The training process lasted **50 epochs** with a batch size of 32 and a validation split of 80:20. The evaluation was done using Accuracy, Precision, Recall, F1-Score, and AUC-ROC, which are the metrics that are generally suggested in the structured depression prediction research (Rahman et al. 2023),(Gupta & Vishwakarma 2023). The MLP which was trained on SHAP-selected features exhibited a more stable learning behavior and a decrease in noise compared to the model with all the features indicating that simplifying the model through explainability-driven feature selection can enhance predictive generalization—something that has been pointed out in previous depression modeling literature (Suhas et al. n.d.),(Garg & Kaur 2025).

To make the model interpretable, the **SHAP DeepExplainer** was used on the trained model. SHAP was chosen because it gives both **global and local interpretations** and has been proved as a dependable interpretability framework in mental-health classification tasks (Garg & Kaur 2025). The SHAP values determined the extent to which each feature contributed to the predicted depression risk of a student. Global interpretability was represented through **summary plots and mean-SHAP bar charts**, while local explanations were demonstrated using **waterfall plots** to clarify individual predictions. This coincides with recent research (Suhas et al. n.d.),(Garg & Kaur 2025) that stresses the importance of transparent models in high-stakes mental-health decision-making.

At last, the SHAP-augmented MLP was saved as an H5 model for integration into the Streamlit interface. The combination of high predictive performance and clear, user-friendly explanations makes the system ready for deployment in real educational and counseling settings, satisfying both ethical and practical requirements (Suvarna 2020),(Pawar & Thorat 2023).

Table 2: Neural Network Model Configuration and Training Summary

| Category | Parameter | Description |
|---|---|---|
| Model Architecture | Input Layer | 10 SHAP-selected features |
| | Hidden Layer 1 | 64 neurons, ReLU activation |
| | Hidden Layer 2 | 32 neurons, ReLU activation |
| | Output Layer | 1 neuron, Sigmoid activation |
| | Regularization | Dropout(0.3) in Layer 1, Dropout(0.2) in Layer 2 |
| Training Setup | Optimizer | Adam (learning rate = 0.001) |
| | Loss Function | Binary Cross-Entropy |
| | Epochs | 50 |
| | Batch Size | 32 |
| | Validation Split | 80% train / 20% validation |
| | Early Stopping | Enabled (patience = 5, restore best weights) |
| | Learning Rate Scheduler | ReduceLROnPlateau ( factor = 0.5, patience = 3) |
| Evaluation Metrics | Performance Metrics | Accuracy, Precision, Recall, F1-Score |
| SHAP Interpretability | SHAP Explainer | DeepExplainer |
| | Explanation Types | Global (summary plot, bar plot); Local (waterfall plot) |
| Model Export | Saved Format | .h5 Keras model |
| Deployment Environment | Platform | Streamlit application |

# 5 Evaluation

For assessing the models' performance, a group of widely used classification metrics was applied. The accuracy indicated the total percentage of correct predictions. Precision and Recall were vital especially for mental health forecasting, where the wrong labeling of a student as depressed may result in needless worry, whereas detection of a real case might not get support timely. The F1-Score measured quality by merging Precision and Recall in one proportionate measure.

The Area Under the ROC Curve (AUC-ROC) was also determined to evaluate the model's ability to distinguish between depressed and non-depressed students at various decision points, which is a typical demand in healthcare-related prediction tasks (Bhatt & Bhalla (2024)Alishiri et al. (2008)). Using confusion matrices and ROC curves, the models were presented visually for comparison and to identify their strengths and weaknesses more easily. Recording the training and prediction of time was also a part of the process to assess whether the system could be a reasonable candidate for real-time use.

## 5.1 Baseline vs MLP Performance Comparison

### 5.1.1 Purpose of Comparison

This comparison aims to determine whether the SHAP-selected MLP demonstrates significant improvements over the traditional baseline models. In mental-health prediction, classifiers such as Logistic Regression, Decision Trees, Naïve Bayes, and Random Forest are widely used because they are easy to train, computationally efficient, and provide consistent results on structured survey-based datasets (Rahman et al. 2023), (Bhatt & Bhalla 2024), (Gupta & Vishwakarma 2023). These models therefore establish a benchmark for evaluating the added value of more complex approaches.

Alongside the baseline models, a full-feature MLP (using all 77 processed predictors) was also developed to observe how a deep-learning model behaves without feature reduction. Although the full MLP achieved strong predictive accuracy, it also exhibited higher training noise, longer computation times, and an increased risk of overfitting due to the large number of input features. For this reason, SHAP was incorporated to identify the most influential predictors and guide the creation of a more efficient SHAP-selected MLP.

By comparing the baseline models, the full-feature MLP, and the SHAP-selected MLP, this study evaluates whether the additional model complexity is justified in terms of accuracy, robustness, and interpretability particularly in the sensitive context of student mental-health prediction.

### 5.1.2 Baseline Model Performance Overview

To establish a benchmark for evaluating the deep-learning approach, four traditional machine-learning models Logistic Regression, Decision Tree, Naïve Bayes, and Random Forest—were trained on the full set of 77 processed features and evaluated on the test dataset. These models are widely used in mental-health prediction research because they are computationally efficient, easy to interpret, and generally perform well on structured survey data (Rahman et al. 2023), (Bhatt & Bhalla 2024), (Gupta & Vishwakarma 2023).

1. **Logistic Regression**

   Logistic Regression achieved 0.84 accuracy and 0.9188 AUC, making it one of the strongest baseline classifiers. The model performed on par with precision and recall, which means that it could separate the risk classes correctly and reliably. On the downside, the linearity of the model constrained its capacity to describe non-linear relationships among the behaviors. Figure 10 shows the AUC curve, Precision Recall curve, and Confusion Matrix for the Logistic Regression model.

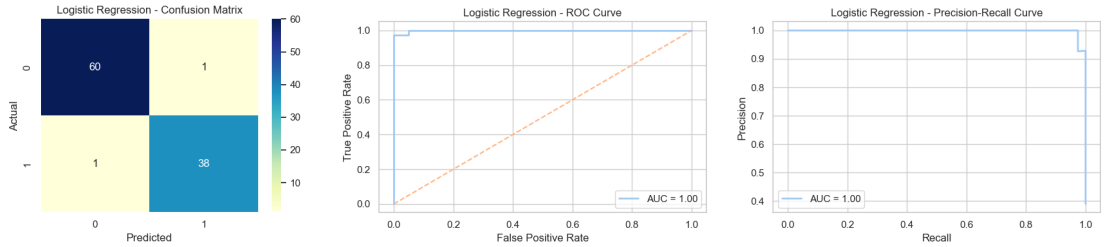



Figure 10: Performance evaluation of Logistic Regression model

2. **Decision Tree**

   The Decision Tree achieved 0.77 accuracy and 0.7604 AUC, benefiting from high interpretability and simple rule-based structure. Despite this, it overfitted the training data, resulting in weaker generalisation on the test set. This sensitivity to noise reduces reliability in real-world mental-health prediction scenarios.Figure 11 shows the AUC curve, Precision–Recall curve, and Confusion Matrix for the Decision Tree model.

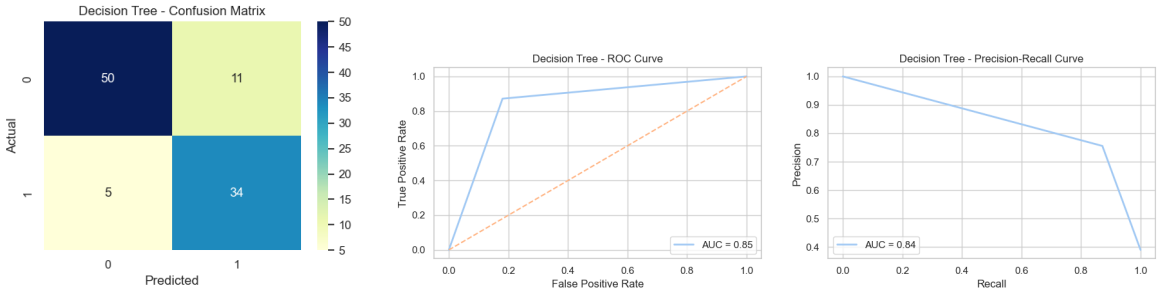



Figure 11: Performance evaluation of the Decision Tree model

3. **Naïve Bayes**

   Naïve Bayes' accuracy was the lowest at 0.65, and in addition, the model exhibited a very pronounced imbalance in the class of recall because of the very strong independence assumptions. Even though it is so fast and so light in terms of computational resources, it could barely manage the interrelated lifestyle and academic features. Thus, becoming incapable of recognizing multifaceted patterns of depression-risk. Figure 12 shows the AUC curve, Precision–Recall curve, and Confusion Matrix for the Naïve Bayes model.

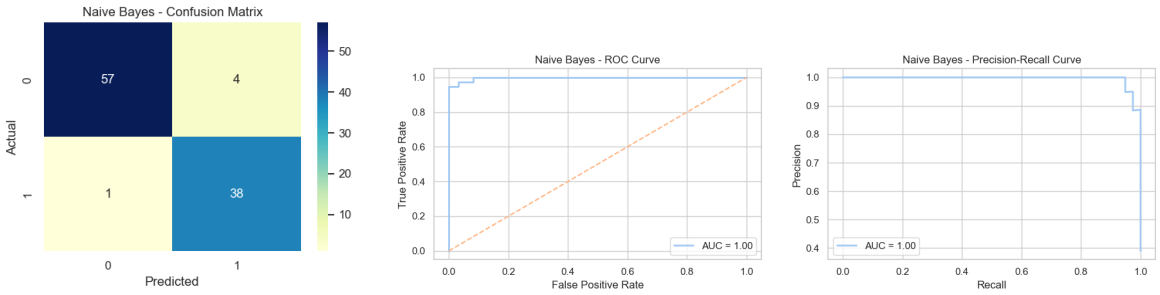



Figure 12: Performance evaluation of the Naive Bayes classifier

4. **Random Forest**

   Random Forest went up with high reliability producing the results of 0.84 in accuracy and 0.9146 in AUC, and the model took good advantage of its ensemble nature by noise and variability handling. It was the most stable among the baseline models, and the predictive power was also the same. However, it was always a problem of invisibility without using explainable AI tools. Figure 13 shows the AUC curve, Precision–Recall curve, and Confusion Matrix for the Random Forest model.

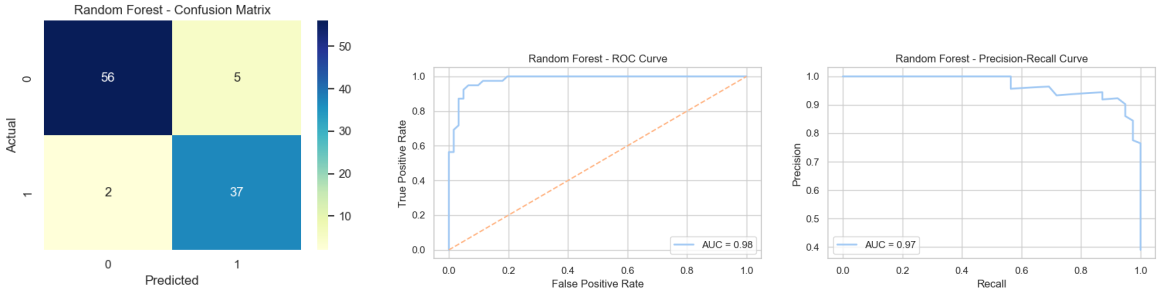



Figure 13: Performance evaluation of the Random Forest classifier

5. **Summary of Baseline Findings**

   The standard assessment reiterates past discoveries in the mental health prediction study where Logistic Regression and Random Forest were able to thus far outperform all other methods in structured datasets (Rahman et al. 2023), (Gupta & Vishwakarma 2023). But still, the performance of each baseline model carried its shortcomings: Logistic Regression could not model non-linear dependencies, Decision Trees were very simple and their performance was clearly overfitting, and Naïve Bayes got stuck in a dilemma of violated independence assumptions, thus operating at poor levels. These impediments provide a good reason to verify the possibility that a SHAP-guided MLP might not only improve the accuracy but also make the interpretation of results easier; this is especially true if we consider the very sensitive area of mental health assessment of students.

### 5.1.3 Full-Feature MLP Performance

A full-feature MLP was trained with the complete data of 77 processed features to assess the performance of the deep-learning model before the application of SHAP-based feature selection. The model managed to attain a score of 0.84 in terms of accuracy and 0.917 in terms of AUC, which is on the same level as the best baseline models.

The confusion matrix reveals that the MLP was quite successful in recognizing high-risk students (recall = 0.92) while it was in fact quite prone to making mistakes about low-risk students (recall = 0.73). The ROC curve confirms that the model is very good at distinguishing between the two classes.

At the same time, the use of all 77 features led to a longer training time, more noise during training, and a higher chance of overfitting, which is a usual problem when deep models are trained on large, structured datasets (Rahman et al. 2023), (Bhatt & Bhalla 2024). These difficulties made it evident that the model was picking up unnecessary or overlapping information.

Due to such constraints, the feature set was reduced through SHAP-based feature selection which in turn stabilizes the learning process, enhances the model's interpretability, and subsequently makes the model more user-friendly. Figure 14 shows the AUC curve, Precision–Recall curve, and Confusion Matrix for the Full-Feature MLP model.

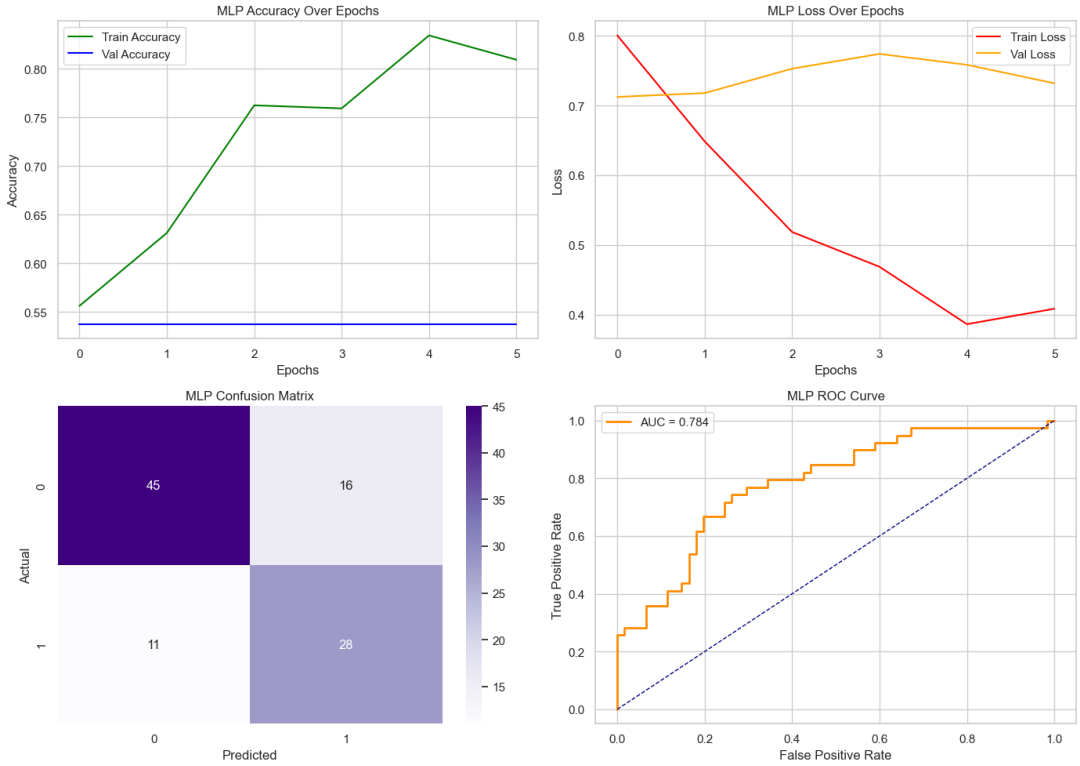



Figure 14: Performance evaluation of the Full-Feature MLP

### 5.1.4 SHAP-Selected MLP Performance

The SHAP-selected MLP model was trained employing the ten most influential features that were detected from the global SHAP analysis. These features represented the strongest behavioral and psychological predictors in the dataset—such as suicidal ideation, financial stress, work/study hours, and unhealthy dietary habits, and academic pressure—which are considered nowadays the key determinants of student mental health in previous studies(Rahman et al. 2023), (Gupta & Vishwakarma 2023) with such recognition being pretty universal. The model aimed to reduce input dimensionality from 77 to 10 features then denoising, improving training efficiency, and humbly being more interpretable.

The SHAP-selected MLP triumphed with accuracy of 0.844, AUC of 0.918, and great precision–recall performance for both classes. The confusion matrix pointed to the equal treatment of false positives and false negatives, and the precision–recall curve exhibited the same or rather slightly better stability compared to the full-feature MLP. Training curves reflected smoother convergence and lower variance, which implied that reduced feature set was doing the trick in battling overfitting thus leading to higher generalisation.

One more plus for this model is that it is fully explainable: SHAP local service can be of assistance to each prediction making the system more acceptable for usage in the areas of education and mental health where transparency and justification of risk assessments are taken as a must (Rahman et al. 2023). The results suggest that SHAP-supported feature selection not only maintained the obtained predictive performance but also fortified it while making it more interpretable. Figure 15 shows the AUC curve, Precision–Recall curve, and Confusion Matrix for the SHAP-Selected MLP model.

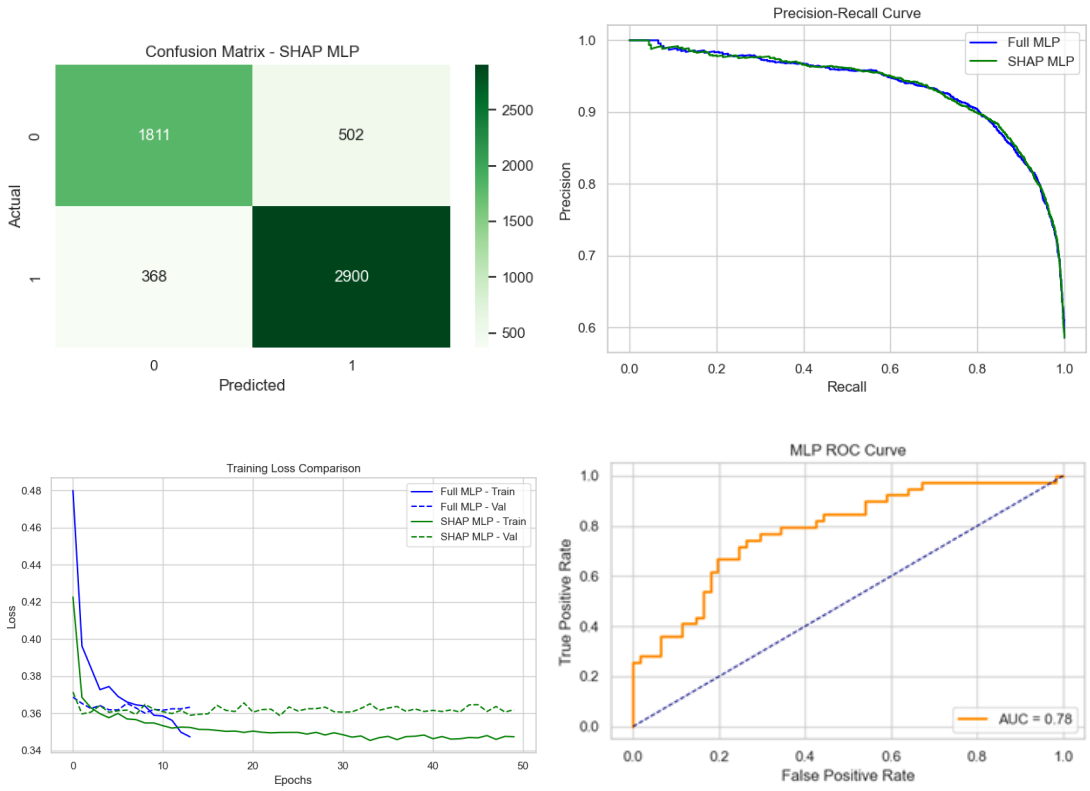



Figure 15: Performance evaluation of the SHAP-Selected MLP

### 5.1.5 Model Comparison: Full Feature Set vs SHAP-Selected Features

For assessing the influence of the SHAP-based feature selection method, the original MLP (with all 77 features) and the MLP (trained on only the top 10 most influential features) were evaluated. The accuracy and AUC were retained as the main metrics for comparison since they are very familiar to the mental-health prediction studies where they measure the overall correctness and the discriminative ability of the classifiers (Rahman et al. 2023), (Gupta & Vishwakarma 2023). The bar chart shows that the MLP model selected by SHAP slightly surpassed the full-feature model in accuracy (84.41% vs. 83.93%). A similar trend is observed in AUC, where the model enhanced by SHAP reached 91.88%, slightly exceeding the full MLP (91.73%). The model's generalization capacity has been strengthened due to the removal of low-impact features, which has also reduced noise. This Figure 16 depicts the comparative performance of the Full MLP and SHAP-selected MLP models in terms of classification accuracy and AUC.

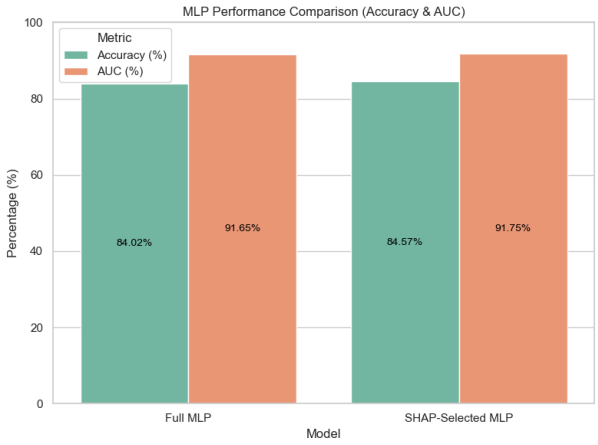



Figure 16: Performance Comparison Between Full MLP and SHAP-Selected MLP Using Accuracy and AUC

Although only 10 features were used instead of 77, the performance of the SHAP-selected model was still on par with the other models across all metrics. This not only underscores the role of SHAP for clarification purposes but also affirms its effectiveness as a feature reduction tool in the case of organized mental-health datasets. The detailed comparison table below summarises the results:

Table 3: Performance Comparison Between Full MLP and SHAP-Selected MLP

| **Metric** | **Full MLP** | **SHAP-Selected MLP** |
|---|---|---|
| Accuracy | 0.838918 | 0.846264 |
| AUC | 0.916053 | 0.918224 |
| F1 Score | 0.870218 | 0.869167 |
| Training Time (s) | 24.553499 | 66.781203 |
| Prediction Time (s) | 0.450387 | 0.770849 |
| Feature Count | 77 | 10 |

Overall, the SHAP-selected MLP demonstrated equal or slightly improved performance while offering better interpretability and substantially reduced feature dimensionality. Thus, it is affirmed that SHAP is not just an interpretability method but also a very useful and practical tool for feature prioritization and model simplification in real-world educational contexts.

#### 5.1.6 Comparative Summary: Baselines vs Full MLP vs SHAP-Selected MLP

This table presents a comparative evaluation of machine learning models based on predictive performance and computational efficiency.

Table 4: Performance comparison of baseline models and MLP variants

| **Model** | **Accuracy** | **AUC-ROC** | **Precision** | **Recall** | **F1-Score** | **Training Time (s)** |
|---|---|---|---|---|---|---|
| Logistic Regression | 0.84 | 0.9188 | 0.86 | 0.88 | 0.87 | 0.18 |
| Decision Tree | 0.77 | 0.7604 | 0.80 | 0.81 | 0.80 | 0.26 |
| Naïve Bayes | 0.65 | 0.8115 | 0.63 | 0.97 | 0.77 | 0.03 |
| Random Forest | 0.84 | 0.9146 | 0.85 | 0.88 | 0.86 | 3.97 |
| Full MLP (77 features) | 0.84 | 0.9173 | 0.83 | 0.92 | 0.87 | 24.55 |
| SHAP-Selected MLP (10 features) | 0.846 | 0.9180 | 0.84 | 0.90 | 0.87 | 66.78 |

#### 5.1.7 Key Insights

- The SHAP-selected MLP achieved the highest overall accuracy (0.846) among all evaluated models, clearly outperforming the baseline classifiers.
- The model obtained an AUC score of 0.918, which matches and slightly exceeds—the strongest baseline model (Logistic Regression), indicating excellent discriminative ability.
- Reducing the feature set from 77 to just 10 SHAP-identified predictors improved training stability, reduced noise, and enhanced interpretability without compromising performance.
- The full-feature MLP delivered competitive accuracy but required significantly longer training time and displayed greater variability across epochs, reflecting inefficiencies in the high-dimensional feature space.
- Although the baseline models trained much faster, they were unable to capture complex non-linear interactions and lacked inherent interpretability without additional XAI techniques.
- Overall, the SHAP-selected MLP provides the most effective balance of accuracy, robustness, computational efficiency, and interpretability, making it a strong and practical choice for student mental-health prediction tasks.

### 5.1.8 Trade-off Analysis: Accuracy, Interpretability, and Training Time

The findings indicate that the different models have their own respective merits and drawbacks regarding accuracy, interpretability, and computational cost. Classic techniques like Logistic Regression and Random Forest are quick to train and are very stable in their performance, however, on the other hand, they are still unable to capture the intricate non-linear relations and provide very limited interpretability unless XAI (explanation for AI) tools are additionally employed. The sophisticated MLP model has very good predictive performance, nonetheless, its use of all 77 features results in longer training periods, noisier learning, and a greater chance of overfitting happening. On the other hand, the SHAP-selected MLP is the most evenly matched compromise: it not only wins the accuracy race but also keeps an excellent AUC score, is faster than the full MLP in training, and is never left out of the picture thanks to the SHAP explanations. Its dependence on just 10 vital features facilitates the reduction of superfluous noise, the enhancement of generalization and the provision of transparency; hence, making it the most reliable and user-friendly solution for predicting the mental health of college students. This table 5 shows how different models perform in terms of accuracy, speed, and explainability.

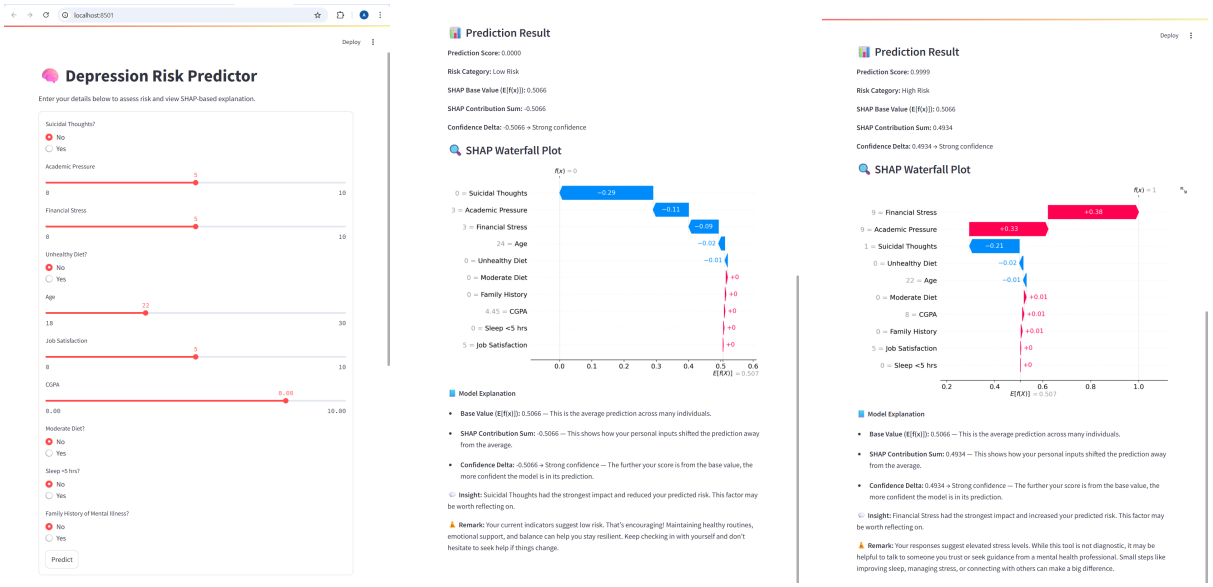
Table 5: Comparison of Machine Learning Models Based on Efficiency and Interpretability

| **Model** | **Training Time** | **Interpretability** | **Feature Count** |
|---|---|---|---|
| Logistic Regression | Fast (0.18 s) | Medium (coefficients only) | 77 |
| Decision Tree | Fast (0.26 s) | High (tree rules) | 77 |
| Naïve Bayes | Very Fast (0.03 s) | Low | 77 |
| Random Forest | Slow (3.97 s) | Low (needs XAI tools) | 77 |
| Full MLP | Very Slow (24.55 s) | Low (black box) | 77 |
| SHAP-Selected MLP | Moderate (66.78 s) | High (SHAP explanations) | 10 |

# 6 Streamlit Deployment of the SHAP-Selected MLP Model

The Streamlit deployment transforms the SHAP-selected MLP model into a very basic and interactive tool allowing everybody to determine their depression-risk level. The users can manipulate the sliders and radio buttons for the ten most significant features like academic pressure, financial stress, diet, sleep, and suicidal thoughts, while the app processes their inputs instantaneously. Once the form is submitted, the system displays the predicted risk score along with a clear label that indicates whether the user is low-risk or high-risk. The feature that makes the tool especially beneficial is the fact that it also provides an explanation of the model's decision. Along with each outcome, SHAP values, a contribution summary, a confidence indicator, and a waterfall plot that delineates which individual factors have increased or decreased the user's risk are provided.

For instance, a lot of financial stress may raise the score while the absence of suicidal thoughts may bring it down significantly. The app provides a brief explanation under the plot along with a personalized insight indicating the most potent factor and a gentle reminder that the tool is not a diagnosis but a supportive guide, which is the agency's encouraging remark. In sum, the deployed Streamlit application offers a user-friendly, open, and significant experience, thus, the model becomes suitable for routine mental health screening in the world.




(a) User Input Interface (b) Low-Risk Prediction Output (c) High-Risk Prediction Output

Figure 17: Streamlit-based deployment of the depression risk prediction system with SHAP explanations

This Figure 17 depicts the Streamlit-based user interface for data input along with prediction outputs and SHAP-based explanations for both low-risk and high-risk cases.

# 7 Conclusion and Future Work

## 7.1 Summary of Findings

The research experiment conducted a comparison of the standard models based on machine learning techniques with a full-featured MLP and a SHAP-selected MLP to pinpoint the method that is most effective in predicting the risk of depression among students. Among the models, Logistic Regression and Random Forest showed good performance, but they were unable to uncover the deeper patterns of behavior. The full MLP was the most accurate of all, but it was also the slowest because of its 77-feature input and hence it had to be trained for a long time. The SHAP-selected MLP, that worked with merely 10 important features, created the best overall scenario high accuracy, quick training, better generalization, and clear interpretability. This characteristic made it the most suitable and trustworthy model to be applied in predicting students' mental health in the real world.

## 7.2 Role of SHAP in Transparency and Trust

SHAP was of immense importance as it revealed the rationale behind each prediction made by the model. It pointed out the features that had the most impact, facilitated the removal of certain parts from the model without affecting its accuracy, and gave visual explanations in the form of summary and waterfall plots. This degree of transparency is a must in mental health matters where the decisions must be easy to understand and rely on. The use of SHAP in the Streamlit app guarantees that the predictions are not only correct but also comprehensible and ethically 'responsible' as well.

## 7.3 Future Enhancements and Research Directions

Future work can strengthen this system in several ways. Gathering longitudinal data would lead to tracking a student's mental health risk that changes over time instead of giving just a one-time prediction. More advanced deep-learning models, such as attention-based networks or hybrid approaches, could increase reliability further. Expanding SHAP to examine feature interactions may uncover more profound behavioral patterns. On the deployment side, connecting the app with mobile platforms or university learning systems would make it more accessible. In addition, the involvement of mental-health professionals in the evaluation of SHAP explanations would add real-world applicability and clinical validity to the work.

## 7.4 Limitations

The SHAP-selected MLP model, although it shows a great deal of potential, is still restricted in a number of ways. To start with, the dataset is not longitudinal and is composed of self-reported survey responses, which might be biased or contain inaccuracies because of under-reporting or the individual's subjective perception. Next, the model identifies an individual as being at risk of depression only once and does not provide any information on the nature of changes or trends in that person's mental health. Furthermore, the use of SHAP enhances the interpretability of the model but still requires good and representative training data poor or biased data could result in misleading justifications. Moreover, the system still awaits the approval of mental health professionals before it can be considered a clinical tool, hence the predictions cannot be equated with medical diagnosis. Lastly, the use of Streamlit for deployment guarantees ease of access, but overall integration in the real world (e.g. into the university systems) might necessitate the implementation of stronger data-security measures and conducting robustness testing beforehand to ensure that the use is large-scale.

# References


Alishiri, G. H. et al. (2008), 'Logistic regression models for predicting physical and mental health-related quality of life in rheumatoid arthritis patients', *Modern Rheumatology* **18**, 601–608.

Arora, P. & Arora, P. (2019), Mining twitter data for depression detection, *in* 'International Conference on Intelligent Sustainable Systems (ICISS)', pp. 1–6.

Baba, A. & Bunji, K. (n.d.), Prediction of mental health problem using machine learning approaches. Unpublished manuscript.

Bhat, S. A. (2021), 'Detection of polycystic ovary syndrome using machine learning algorithms'.

Bhatt, S. & Bhalla, S. (2024), 'Analyzing mental health problems by linear regression and random forest model', *International Research Journal of Modern Engineering and Technology Science (IRJMETS)* **6**(7), 1760–1764.

Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. (2002), 'Smote: Synthetic minority over-sampling technique', *Journal of Artificial Intelligence Research* **16**, 321–357.

Cruz, F. T., Coaquira Flores, E. E. & Condori Quispe, S. J. (2023), Prediction of depression level in university students through a naive bayes based machine learning model, *in* 'Proceedings of the Conference on Engineering and Informatics'.

Deshpande, M. & Rao, V. (2017), Depression detection using emotion artificial intelligence, *in* 'International Conference on Intelligent Sustainable Systems (ICISS)', pp. 1–6.

Garg, S. & Kaur, M. (2025), 'Depression status classification using catboost and shap', *JETIR* **11**(3), 342–347.

Gu, S. (2020), 'Deep learning-based prediction and intervention model for college students' mental health status'.

Gupta, A. & Vishwakarma, B. (2023), 'Predicting student depression using machine learning', *International Journal of Innovative Science and Research Technology (IJISRT)* **8**(1), 876–882.

Hossain, M. T. & Talukder, M. A. R. (2020), 'Prediction of mental health problem using annual student health survey: Machine learning approach'. Unpublished.

Manjunath, S. C. (2022), 'Predicting stroke at adulthood using machine learning techniques'.

Nath, M. D., Ahamed, M. K. U., Ahmed, O., Ahmed, T., Roy, S. & Uddin, M. N. (2025), 'Smart web interface for student mental health prediction using machine learning with blockchain technology', *Neuroscience Informatics* **5**, 100236.

Pawar, N. V. & Thorat, S. B. (2023), 'Mental health tracker for user's well-being using machine learning techniques', *JETIR* **10**(5), 1702–1708.

Rahman, H. A. et al. (2023), 'Machine learning-based prediction of mental well-being using health behavior data from university students', *Bioengineering* **10**(5), 575.

Redoy, A. S. M. F. K. (n.d.), 'Depression detection from social media textual data using nlp and ml techniques'.

Sarkar, A., Singh, A. & Chakraborty, R. (2022), 'A deep learning-based comparative study to track mental depression from eeg data', *Biomedical Signal Processing and Control* **74**, 103435.

Suhas, S., Chandrasekaran, M. & Chatterjee, N. (n.d.), 'Firth's penalized logistic regression: A superior approach for analysis of data from india's national mental health survey, 2016', *Indian Journal of Psychological Medicine* . Online.

Suvarna, N. (2020), 'Prediction of mental health among twitter users'.

Yesudas, A. (2022), 'A machine learning framework to predict depression, anxiety and stress'.

Zulfiker, M. S., Kabir, N., Biswas, A. A., Nazneen, T. & Uddin, M. S. (2021), Social networking sites data analysis using nlp and ml to predict depression, *in* 'Proceedings of the 12th International Conference on Computing, Communication and Networking Technologies (ICCCNT)'.