# Mind Matters: Harnessing Machine Learning to Predict Psychiatric Diagnoses in University Students with Real-World Data

Jake Watts, Alex LeBouthillier, Matteo Passalent, Noah Fedosoff, Vikas Movva, Griffin Lind,
Tamal Chakroborty, Yang Liu*
Wilfrid Laurier University
{watt7490, lebo8020, pass7410, fedo0350, movv7230, lind2052, chak0440}@mylaurier.ca
yangliu@wlu.ca

*Abstract*—**By March 2020, the COVID-19 pandemic drastically altered daily life, with repercussions for mental health. This study investigates the use of demographic, lifestyle, and psychometric data to predict psychiatric diagnoses among undergraduate students. Leveraging survey responses, we explore machine-learning approaches to enable scalable, cost-effective mental health interventions. Results demonstrate the efficacy of various classifiers, with the Multi-layer Perceptron Model emerging as the most effective.**

*Index Terms*—**Multi-layer Perceptron (MLP), Cognitive Affective Scale of Mindfulness Revised (CAMS-R), psychiatric diagnoses, Generalized Anxiety Disorder (GAD).**

## I. Introduction

By March 2020, the COVID-19 virus had become a global threat to public health. In response, many institutions rushed to mitigate its effects by restricting in-person contact through measures such as working and studying from home. The idea was to limit contact to prevent the virus from spreading as quickly, which would help “flatten the curve” of infected persons requiring medical treatment, reducing the strain on healthcare systems. Though well-intentioned, these measures disrupted people’s everyday lives and had negative effects on their mental health. This context sets the stage for the problem being investigated in this study.

In particular, we investigate whether an individual’s demographic circumstances, lifestyle, and mental state during the pandemic can predict the presence of a diagnosed psychiatric disorder. To accomplish this, we leveraged a dataset from Paterson and Reeves [1], which was collected during the fall of 2020 through an online survey. The survey covered different sections including demographics, mental health, changes in mental health during the pandemic, self-care habits, and academic hobbies [12]. The researchers created an online survey using the Qualtrics platform and solicited responses from participants who were self-reported as university students registered in a four-year undergraduate program in Canada. This comprehensive survey included a wide range of questions designed to provide a detailed overview of each participant’s life and mental health status at that time.

The dataset provided us with the opportunity to train machine learning models that can predict psychiatric diagnoses among students during an unprecedented global pandemic or global emergencies. Specifically, universities could utilize such a model to proactively identify students at risk of experiencing poor mental health outcomes and take preventive measures to mitigate these adverse effects. The problem we tackle in this paper will be predicting whether university students have a psychiatric diagnosis in the context of a large-scale upheaval.

We investigated the use of a number of models, eventually settling on a battery of four classifiers including Support Vector Machines, Multi-layer Perceptrons, Gradient Boosting, and Naive Bayes. Additionally, we examined the psychometric scales present in the data and demonstrated how these scales relate to the psychological impact on students during the global pandemic.

There is a significant amount of research [2]–[5] surrounding machine learning and mental health. However, our work stands out due to its potential applications and the types of data we used for model training. A recent examination of current multi-modal methods [4] shows a large body of work surrounding the use of machine learning models to interpret multimodal data, such as video, images and EEG signals. Further exploration of the literature indicates that recent studies are applying large language models to diagnose specific disorders, including depression [5].

In our research, we trained models based on categorical survey responses to various psychometric scales. This approach allows scalable and non-invasive data collection, such as online surveys, which healthcare providers can use to identify individuals who may benefit from personalized care. Fong et al. [6] introduced an MLP model to predict mental health trends, particularly anxiety and depression, using a dataset collected during the COVID-19 pandemic in a diverse cohort of U.S. adults. However, Our research differs by targeting a general assessment of mental health disorders rather than specific conditions. Our work contributes to available literature by providing an easily scaled and low-cost method for distinguishing individuals with mental health conditions. It does not rely on multi-modal data, which can be costly and time-consuming to collect. It applies to various mental health issues rather than being limited to a specific condition.

## II. Data Exploration and Preprocessing

The dataset derived from survey responses [1], included several established psychometric scales and various demographic and lifestyle features. Psychometric scales provided aggregated scores representing constructs such as emotional regulation and mindfulness, while demographic features included age, sex, ethnicity, and academic-related data [8]. However, there are some inconsistencies in the dataset, and to prepare the data for machine learning models, certain attributes needed to be converted into specific formats.

### A. Feature Engineering

The survey included a "catch question" meant to weed out respondents who were not paying attention to their answers. We used the values of this feature to drop all records that had failed the catch question. Since the dataset was collected online and participants self-reported their eligibility, this was one of the most effective methods available to ensure the accuracy of our data. After this step, we removed the feature containing the catch question responses. The survey included a set of questions assessing the importance $(Hobbies\_Imp)$ and the amount of time spent $(Hobbies\_Time)$ on various hobbies. These were divided into two broad categories: Recreational and Academic. Each category contained four specific hobbies, with corresponding columns in the dataset. To simplify and consolidate this information, we decided to aggregate them into new, more informative, features. An academic score and a recreational score were calculated for each respondent. The recreational score was the sum of $Hobbies\_Imp$ fields 1 to 4 and $Hobbies\_Time$ fields 1 to 4, while the academic score was the sum of the remaining hobbies fields (5-8). Each record now had a number that represented the combined time and importance placed on hobbies of each type. Following this, we dropped the sixteen original hobbies from the dataset. Additionally, we dropped ten more unrelated attributes from the study, such as $StartDate$, $EndDate$, $Eligibility$, $Ethnicity\_text$, $Diagnosis\_text$, $Province$, $Degree$, and $Volunteering$.

### B. Aggregating Psychometric Scales

Psychometric scales are structured instruments designed to measure psychological constructs such as mental health, emotional well-being, or cognitive processes. These scales typically consist of a series of statements or questions that respondents rate or answer based on their experiences, feelings, or behaviours. The goal is to quantify abstract psychological concepts, enabling researchers to assess trends, diagnose conditions, or evaluate interventions [10], [11]. The discussion in this paper focuses on the three scales because of their interesting findings: GAD-2 (Generalized Anxiety Disorder) [14], CAMS-R (Cognitive Affective Scale of Mindfulness Revised) [9], and DERS-16 (Difficulties in Emotion Regulation Scale) [7].

The GAD-2 scale, a brief screener for generalized anxiety disorder, measures symptoms like excessive worry and inability to control anxiety, with scores ranging from 0 to 6. As shown in Fig. 1, pre-pandemic scores were concentrated at the lower end, reflecting mild anxiety levels, whereas post-pandemic scores showed a broader distribution with more participants scoring in the higher range. This shift highlights a clear increase in anxiety levels following the COVID-19 outbreak, consistent with global trends during periods of large-scale upheaval.

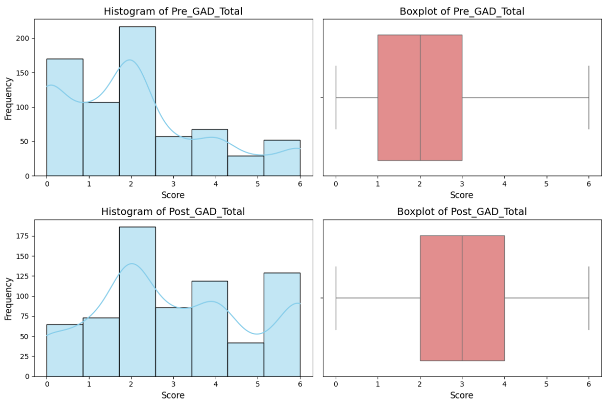


Fig. 1. Visualizing the GAD-2 Scale

The CAMS-R scale assesses mindfulness across four dimensions: attention, awareness, focus, and acceptance (Feldman). Scores range from 12 to 48, and participants generally reported moderate-to-high levels of mindfulness, with scores clustering in the 30–32 range. This result in Fig. 2 highlights the importance of mindfulness in coping with stress and emotional challenges, as mindfulness is strongly associated with improved emotional regulation and mental well-being. A small subset of participants scored exceptionally high on mindfulness, which could indicate the use of mindfulness practices as a coping mechanism during challenging times.

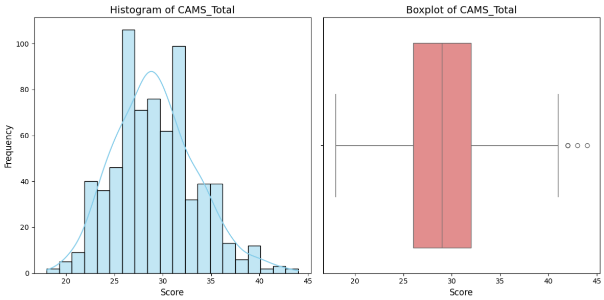


Fig. 2. Visualizing the CAMS-R Scale

The DERS-16 scale evaluates five aspects of emotional regulation: lack of emotional clarity, nonacceptance of negative emotions, impulse control difficulties, limited access to effective regulation strategies, and goal-directed behaviour challenges. Scores range from 16 to 80, with higher scores indicating greater emotional regulation difficulties. It is clear that, in Fig. 3, most participants scored in the moderate range, suggesting a prevalence of moderate difficulties in emotion regulation, while a smaller group reported severe challenges.

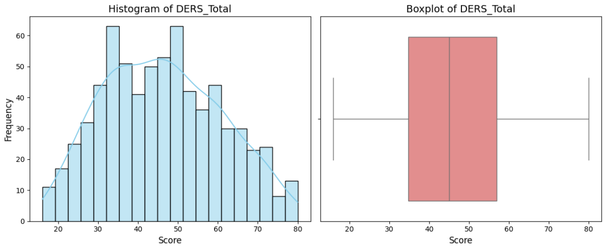


Fig. 3. Visualizing the DERS-16 Scale

### C. *Handling Imbalance Data*

To address the issue of class imbalance in the Diagnosis feature (there were significantly more instances of one level than the other), we applied SMOTE (Synthetic Minority Oversampling Technique) [13] after splitting the dataset into training and testing sets. SMOTE works by creating synthetic samples for the minority class by interpolating between existing minority class instances. This ensures that the training dataset has an equal representation of both classes, which is particularly important for models that may otherwise bias predictions toward the majority class. The application of SMOTE helps improve precision and recall scores by enabling models to better learn from the minority class while still retaining a realistic distribution in the test set for evaluation.

## III. Model Training and Metrics

Given the complexity of the underlying relationships and the relatively small sample size, a variety of machine-learning algorithms were explored. These included Support Vector Machines, Multi-layer Perceptron, Gradient-Boosting Decision Tree, and Naive Bayes classifier.

The rationale for employing multiple modelling techniques arose from the need to identify the approach best suited to capturing subtle patterns in the data. For instance, preliminary analyses suggested that linear correlations among features and the target were low, indicating that more flexible models would potentially yield better predictive performance. At the same time, the modest size of the dataset necessitated the use of methods known to perform reliably in data-constrained environments. The implementation of the models made available in Github public repository [15].

### A. *Gradient Boosting*

The Gradient Boosting (GB) classifier was implemented using decision trees as base learners. By iteratively fitting new trees to the residual errors of the previous ensemble, the model could increasingly refine its predictions. Parameter tuning focused on controlling model complexity through parameters such as $max\_depth$ and employing sub-sampling to introduce randomness and improve generalization. This iterative refinement proved advantageous for handling mixed variable types and for extracting feature importance measures, which aided interpretability. The ability to rank predictors offered insights into which survey questions or demographic factors carried the most predictive weight. The top 10 indicators found are illustrated in Fig. 4.

### B. *Naive Bayes*

The categorical Naive Bayes (NB) model offered a contrast to Gradient Boosting. By making a strong conditional independence assumption among predictors, this probabilistic method could be implemented quickly and interpreted straightforwardly. Before applying the model, continuous features were discretized using uniform binning. This preprocessing step transformed the continuous variables into categorical bins, enabling the model to handle the entire feature set in a unified framework. While the independence assumption is not always realistic, this method served as a baseline and a check against more complex solutions. Its relative computational simplicity and direct probabilistic outputs made it valuable for rapid experimentation and sensitivity analysis related to feature transformations.

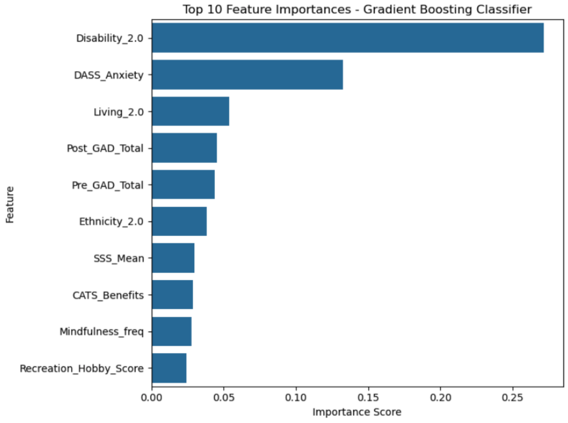


Fig. 4. Visualizing top 10 important features

### C. *Support Vector Machines*

The Support Vector Machine (SVM) model was introduced to leverage its utility in handling complex high-dimensional feature spaces, making it a suitable candidate for the reduced and transformed dataset emerging from our extensive feature engineering steps. Given that the demographic, lifestyle, and aggregated psychometric features did not exhibit strong linear correlations with the target, SVMs provided an opportunity to capture non-linear patterns through the use of kernel functions. Prior to training, continuous features were range normalized to ensure that all predictors contributed proportionally to the model's decision boundary. Hyperparameter tuning focused on selecting appropriate kernels (e.g., linear vs. RBF) and regularization parameters (C, gamma), balancing the SVM's ability to fit nuanced decision surfaces against the risk of overfitting in a relatively data-constrained environment. By iteratively refining these parameters, the SVM model aimed to delineate the subtle relationships between the aggregated psychometric scales, consolidated hobby scores, and demographic/lifestyle variables that could collectively predict mental health diagnoses.

### D. *Multi-layer Perceptrons*

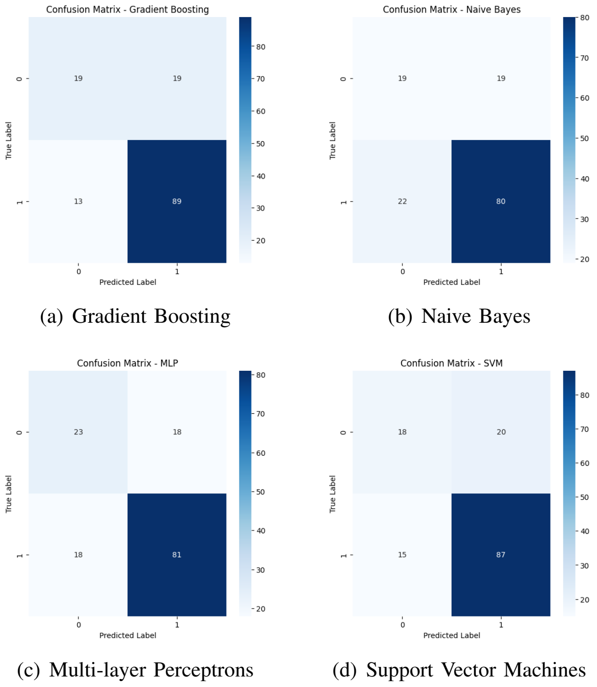
A neural network model was also developed in the form of a three-layer MLP. The architecture comprised an input layer corresponding to the 55 descriptive features, a single hidden layer of moderate size (16 neurons), and a two-neuron output layer for binary classification. A Non-linear activation function ReLU was employed to capture complex interactions that linear models or simple decision structures might miss. To counteract the tendency of neural networks to overfit when data is limited, regularization techniques were integrated into the training process. Specifically, L2 weight decay was applied, and a 50% dropout layer was introduced in the hidden layer. This encouraged the network to rely on more robust patterns rather than memorizing specific samples or feature subsets. The training process also incorporated hyperparameter tuning and validation metrics to ensure generalization. Although more computationally involved, the MLP offered a flexible framework for uncovering intricate relationships in the data.

## IV. Model Evaluation and Comparison

To evaluate and compare the performance of these four different models, we made use of confusion matrices, ROC curves, and precision-recall curves.

### A. *Confusion Matrix*

Confusion matrices in Fig. 5 provide insight into the models' classification abilities by presenting the counts of true positives, true negatives, false positives, and false negatives. The GB classifier showed a high true positive rate with fewer false negative cases, indicating good detection of positive cases. True negative counts were also high, confirming a low false positive rate. NB, while maintaining a balanced true positive and false positive rate, had a slightly higher false negative rate than gradient boosting, pointing to some misclassification of positive cases. The MLP demonstrated competitive true positive and true negative rates, but some potential overfitting resulted in slightly higher false positive rates compared to gradient boosting. The SVM displayed moderate performance with balanced true positive and true negative rates but struggled with false negatives, indicating room for improvement in detecting positive cases.



(a) Gradient Boosting

(b) Naive Bayes

(c) Multi-layer Perceptrons

(d) Support Vector Machines

Fig. 5. Confusion Matrix Evaluation

### B. *ROC Curve*

ROC curves illustrate the trade-off between sensitivity and specificity. As shown in Fig. 6, the MLP classifier achieved the highest area under curve (AUC) score of 0.77, displaying excellent discrimination between classes. The NB classifier followed with a score of 0.73, demonstrating good but slightly weaker performance compared to gradient boosting. The SVM recorded the lowest score of 0.69, showing adequate but less reliable classification performance compared to the other models.

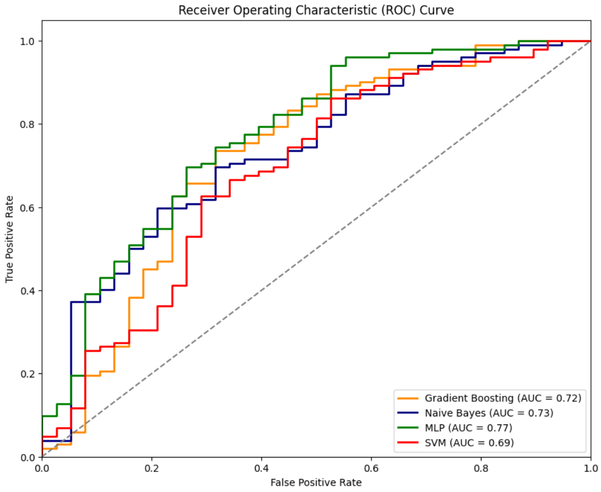


Fig. 6. Visualizing the ROC Curve

### C. *Precision VS. Recall*

Precision-recall curves visualize the balance between precision and recall, which is particularly useful in imbalanced datasets. As indicated in Fig. 7, the MLP classifier consistently maintained a high precision across varying recall levels, confirming its reliability in predicting positive cases. The NB classifier showed good recall but slightly lower precision at higher recall levels, indicating occasional false positives. The SVM's precision and recall were moderate, with noticeable dips in precision at higher recall levels, reflecting its challenge in handling imbalanced data effectively.

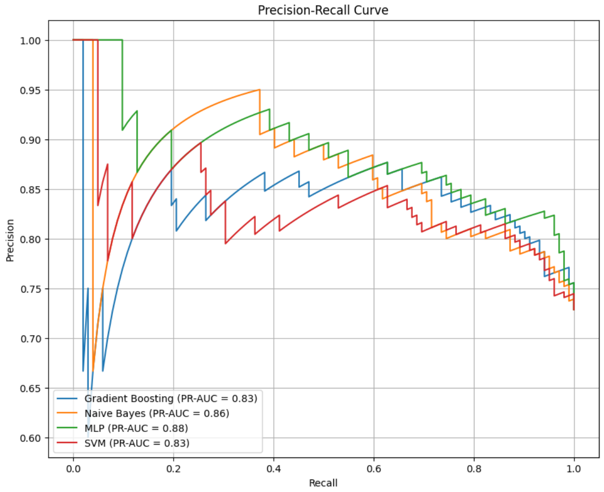


Fig. 7. Visualizing Precision vs. Recall

Based on the evaluation metrics, we found that the MLP classifier emerged as the most effective model, demonstrating superior performance across the confusion matrix metrics, ROC curve AUC, and precision-recall curve. The NB followed closely, showcasing strong capabilities in modelling complex relationships and handling imbalanced datasets. The SVM was the least effective model, requiring further optimization to match the performance of other methods.

## V. Results and Conclusion

This paper has investigated the effectiveness of four different machine learning models in predicting psychiatric diagnoses among university students from a dataset consisting of self-reported survey data. Among the models evaluated, the MLP classifier consistently outperformed others, demonstrating superior performance across the evaluation metrics we employed. Additionally, a direct examination of responses to the various psychometric scales included in the survey, such as the DERS-16 (emotional regulation) and CAMS-R (mindfulness) scales, highlighted the impacts of a large-scale upheaval on these aspects of mental health.

In sum, this paper has demonstrated that survey-based data, combined with machine learning, offers a scalable and cost-effective approach to identifying individuals who may have a psychiatric diagnosis, and thus may benefit from mental-health promoting interventions. By leveraging such models, institutions like universities could implement proactive measures, targeting at-risk students before adverse outcomes materialize. The findings also encourage further exploration of psychometric data and feature engineering to improve predictive accuracy. Future research could expand datasets, test the models in real-world settings, and integrate additional variables to enhance generalizability and impact. Ultimately, this study highlights the potential of machine learning to transform mental health assessment and intervention, fostering healthier and more resilient student populations.

## VI. Acknowledgement

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) Individual Discovery Grants.

## References

[1] T. Paterson and J. Reeves, "University Student Mental Health," *Borealis*, 2022. Available: https://doi.org/10.5683/SP3/VEIBVL. [Accessed: Dec. 7, 2024].

[2] G. Lorenzoni, C. Tavares, N. Nascimento, P. Alencar, and D. Cowan, "Assessing ML Classification Algorithms and NLP Techniques for Depression Detection: An Experimental Case Study," *arXiv preprint*, Apr. 2024. Available: https://arxiv.org/abs/2404.04284. [Accessed: Jan. 19, 2025].

[3] S. Kim and S. Kim, "Automatic Prediction of Mortality in Patients with Mental Illness Using Electronic Health Records," *arXiv*, Oct. 2023. Available: https://doi.org/10.48550/arXiv.2310.12121. [Accessed: Dec. 7, 2024].

[4] Z. Al Sahili, I. Patras, and M. Purver, "Multimodal Machine Learning in Mental Health: A Survey of Data, Algorithms, and Challenges," *arXiv*, Jul. 2024. Available: https://doi.org/10.48550/arXiv.2407.16804. [Accessed: Dec. 7, 2024].

[5] C. Tank, S. Pol, V. Katoch, S. Mehta, A. Anand, and R. R. Shah, "Depression Detection and Analysis using Large Language Models on Textual and Audio-Visual Modalities," *arXiv preprint*, Jul. 2024. Available: https://arxiv.org/abs/2407.06125. [Accessed: Jan. 19, 2025].

[6] D. Fong, T. Chu, M. Heflin, X. Gu, and O. Seneviratne, "Predicting Depression and Anxiety: A Multi-Layer Perceptron for Analyzing the Mental Health Impact of COVID-19," *arXiv*, Mar. 2024. Available: https://doi.org/10.48550/arXiv.2403.06033. [Accessed: Dec. 7, 2024].

[7] Bjureberg, J., Ljótsson, B., Tull, M.T. et al. "Development and Validation of a Brief Version of the Difficulties in Emotion Regulation Scale: The DERS-16," *Journal of Psychopathology and Behavioral Assessment*, vol. 38, no. 2, pp. 284–296, 2016. Available: https://doi.org/10.1007/s10862-015-9514-x.

[8] S. Cohen, T. Kamarck, and R. Mermelstein, "A Global Measure of Perceived Stress," *J. Health Soc. Behav.*, vol. 24, no. 4, pp. 385–396, 1983. Available: https://doi.org/10.2307/2136404.

[9] G. Feldman, et al., "Mindfulness and Emotion Regulation: The Development and Initial Validation of the Cognitive and Affective Mindfulness Scale-Revised (CAMS-R)," *Journal of Psychopathology and Behavioral Assessment*, vol. 29, no. 3, pp. 177–190, 2006. Available: https://doi.org/10.1007/s10862-006-9035-8.

[10] G. Godin, "The Godin-Shephard Leisure-Time Physical Activity Questionnaire," *The Health & Fitness Journal of Canada*, vol. 4, no. 1, pp. 18–22, 2011. Available: https://doi.org/10.14288/hfjc.v4i1.82.

[11] K. Kroenke, et al., "The Patient Health Questionnaire-2: Validity of a Two-Item Depression Screener," *Medical Care*, vol. 41, no. 11, pp. 1284–1292, 2003. Available: https://doi.org/10.1097/01.MLR.0000093487.78664.3C.

[12] J. Moses, et al., "When College Students Look After Themselves: Self-Care Practices and Well-Being," *Journal of Student Affairs Research and Practice*, vol. 53, no. 3, pp. 346–359, 2016. Available: https://doi.org/10.1080/19496591.2016.1157488.

[13] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, no. 1, pp. 321–357, Jan. 2002.

[14] J. García-Campayo, E. Zamorano, M. A. Ruiz, et al., "The Assessment of Generalized Anxiety Disorder: Psychometric Validation of the Spanish Version of the Self-Administered GAD-2 Scale in Daily Medical Practice," *Health and Quality of Life Outcomes*, vol. 10, no. 114, 2012. Available: https://doi.org/10.1186/1477-7525-10-114.

[15] Tamal, "Predicting-Psychiatric-Diagnoses," GitHub. Available: https://github.com/tamal3472/Predicting-Psychiatric-Diagnoses. [Accessed: Jan. 19, 2025]