MHP Mental Health Dataset Classifier

Machine Learning Research Report

By: William Boulton III and Matthew Ong

Contents

Abstract

The increasing prevalence of mental health ailments, such as stress, anxiety, and depression among university students has caused heightened concern in healthcare, especially following the COVID-19 pandemic. Mental disorders have been shown to significantly impact educational attainment and quality of life for students of all backgrounds and departments. The objective of this research project was to assist in preventing academic-related mental illness by using machine learning to classify university students according to their depression severity. This was accomplished by building and comparing the predictive performance metrics of two classification models– logistic regression and XGBoost. We then applied an explainability AI framework known as SHAP to identify the most influential predictors on mental health and therefore extract actionable insights from the best-performing model. This work aligns with ongoing research efforts in applying machine learning and interpretability tools to psychiatric and clinical mental-health domains. Thus, this project is coupled with the same limitations and considerations that apply to the entire field of healthcare. Our results demonstrated the expected high comorbidity of depression with both stress and anxiety. The SHAP analysis provided universities with mental health insights that could enable them to tailor workshops, counseling, or interventions for students at risk in order to reduce attrition and improve educational performance.

1 Introduction

1.1 Problem Statement

This project is to assist in ongoing research to combat mental illness in university students by identifying which aspects of a given student's lifestyle or academic status corresponds to a positive diagnosis of academic depression. The prevalence of stress, anxiety, and depression among university students has seen an increase over the past several years, particularly during the outbreak of COVID-19. Coinciding with the pandemic, increasing academic and financial pressures, along with prolonged periods of isolation continue to contribute to a rise in rates of student depression. These psychological states are serious medical issues that can greatly negatively impact academic performance. As a result, there is a need for reliable tools that universities can use to identify students at risk for depression early on and determine which factors or diagnostic criteria that a university should focus on the most or look for within the student population.

1.2 Objective and Methods

The main deliverable of this project is a tool or model that can predict a college or university student's level or risk of depression using a predetermined set of diagnostic criteria, which may include sociodemographic data, academic data, or mental health assessment scores. The requirements of this project were met using supervised machine learning techniques, which have already been applied to various areas of psychological healthcare and have shown high accuracy in recent years in predicting and treating mental disorders. The study presents a classification problem, where the objective is to be able to predict a student's level of depression severity based on a set of predictors. Various machine learning algorithms are available for us to

use for this purpose; we chose to compare the predictive performance of two models– logistic regression and eXtreme Gradient Boosting (XGBoost). Logistic regression, which allows us to predict the output of a certain categorical target variable, is classification using the architecture of a linear model and is commonly used as a baseline in multi-class prediction tasks. XGBoost is a more complex model, which makes predictions based on a sequence of trees that examine nonlinear decision boundaries between features, that we assessed against the logistic regression model as a benchmark. We then chose the best performing of these two models and applied an explainable AI framework to add interpretability to the model's classifications. This allowed us to learn which predictors from our dataset held the most influence towards the model's output.

2 Literature Review

2.1 Academic Mental Health Research

Current academic research on mental health in university students shows that this group experiences a far higher rate of mood and anxiety disorders compared to the rest of the population. One article reviewed 24 individual studies that were performed between 1990 and 2010 about the prevalence of depression in university students. They found that depression rates among undergraduate students were around 30.6% (Ibrahim, et al., 2012, para. 13). Several factors contributed to a higher risk of depression among students. Age and academic year were found to be linked. Students who were younger or in earlier years were at a higher risk for depression. Gender is also a factor as, in the sixteen studies that included gender, female participants had a 5% higher rate of depression on average (Ibrahim, et al., 2012, paras. 16-17). Another article, which focuses on the results of World Mental Health surveys conducted by the World Health Organization found similar rates and causes of depression among students. 20.3%

of college students met the criteria for a 12-month mental disorder as classified by the DSM-IV. However, 83.1% of those cases began before the students entered college (Auerbach, et al., 2016, paras. 12-13). Cases in that range were associated with low academic performance and higher stress, which can make the transition into college much more difficult. The study also found that students with depression or other disorders had a higher chance of dropping out before graduation (Auerbach, et al., 2016, para. 16). The findings of these two articles emphasize how common mental health problems are in undergraduate student populations and also major gaps in treatment and early detection due to the nature of these problems.

2.2 Risks and Limitations of Machine Learning in Healthcare

There are several important risks and considerations of machine learning in healthcare. One crucial concern is data availability and anonymity. Machine learning models need to be trained on data, and large mental health datasets that have been ethically sourced are in short supply. The privacy of patients whose data is stored in these datasets may not be guaranteed due to the current methods of data collection (Iyortsuun, et al., 2023, para. 5). The article emphasizes a need for informed consent and better explanations for what the data will be used for to reassure patients that their data will not be leaked or used maliciously (Iyortsuun, et al., 2023, para. 60). Another major risk that is tied to the challenge of finding sufficient data is model generalization and overfitting. Mental health datasets can be extremely skewed. The percentage of people without a mental disorder may be much higher than the percentage of people with mental disorders. In the group of people with mental disorders, some disorders may be underrepresented and have less data compared to more common disorders. This can cause the model to misdiagnose one disorder as something else because of the imbalance in the training data. Besides obtaining better data, other methods of reducing overfitting must be considered. The

article states that hyperparameter tuning and including independent test data could help alleviate this problem (Iyortsuun, et al., 2023, para. 59). Addressing these issues in data collection and overfitting is critical in making machine learning models that are used for mental health classification more accurate and more trustworthy for clinicians and patients alike.

2.3 SHAP Explainability Values

To make machine learning models more trustworthy, it is important to be able to understand how a model made a specific prediction. However, many models that are currently being used, in healthcare and other areas, are black-box models. Black-box models prevent us from seeing the inner mechanisms and calculations that led to a specific output. This is a major problem when using machine learning for healthcare, as clinicians need to understand how a model reached a decision so it can be verified. Explainable AI techniques have emerged as ways of interpreting these black-box models. One of these techniques is SHAP, which stands for Shapley Explainability Values. SHAP assigns a shapley value to each feature in a model. This shapley value quantifies the amount of influence a feature has on the model's prediction. There is already some precedent in using SHAP in models that calculate mental disorders. One study from Nanjing used SHAP in combination with deep learning neural networks to predict risk of mental disorder outpatient visits based on environmental factors such as air pollutants. Shapley values were computed for each input variable (each pollutant that was detected), and those values indicated that certain pollutants had more of an influence on a person's risk than others (Wang, et al., 2021). This study shows that having both local explanations for individual predictions and global explanations for overall feature importance makes a black-box model more interpretable and trustworthy.

2.4 XGBoost

One of the models we have used here is XGBoost. XGBoost stands for Extreme Gradient Boosting. It performs very well when used for classification tasks, meaning it should work well for mental health analysis. One of the 24 studies that were examined in the article by Iyortsuun, et al. used XGBoost to classify depression. The model produced extremely impressive results, with an accuracy of 97%, precision of 95%, and recall of 99%. Another study that was reviewed in the same article also used XGBoost, but instead of classifying depression, it was attempting to differentiate between people who are neurotypical and people who have ADHD. The results from this model were not as impressive as the previous, but they were still acceptable. It achieved an accuracy of 74% across independent testing. The lower accuracy could be attributed to a smaller sample size (the first study had 11,081 samples and the second had only 360). It may have also been because of how symptoms of ADHD often overlap with the behaviors of other disorders such as ASD and even neurotypical behaviors (Iyortsuun, et al., 2023, para. 22). This makes classifying it more difficult than other disorders. Together, these two studies demonstrate that XGBoost can perform effectively across multiple mental health classification tasks, even when they may be difficult or more complex. Despite its strong performance, we may not be able to trust the results because XGBoost is generally considered to be a black-box model. Mental health classifications require interpretability and justification. The SHAP scoring mentioned previously can be applied to the XGBoost model to provide the interpretability that is needed.

3 Predicting Depression Severity Levels for Students in Bangladesh Universities

3.1 Dataset Description and Visualization

The dataset that we used to train the models is the Mental Health Problems (MHP) dataset for university students. The dataset contains survey responses from 2028 students across the top 15 universities in Bangladesh. The survey contains three diagnostic mental health models, namely GAD-7, PSS-10, and PHQ-9, which were adapted into questionnaires designed for university students. The models ask specific questions pertaining to anxiety, stress, and depression respectively. The intended goals for usage of this dataset as outlined by the original publishing article are "to explore the trajectory of the mental and psychological stressors faced by the university students" and "to explore the socio-demographic drivers that influence different mental health stressors" (Syeed, et al., 2024, para. 1). The same article also explicitly encourages the usage of the dataset for the purpose of building data-driven machine learning models to predict mental health conditions. The data was collected through collaboration with faculty members from the participating universities who distributed and conducted the surveys via Google Forms. The severity levels for anxiety, stress, and depression were determined by converting each questionnaire's total score into categories using predefined thresholds. The target category for our concerns was the responder's depression level, or the final column, using all other predictors as input features. The reason we chose depression as the primary target was because, as aforementioned, while stress and anxiety may be transient for many cases, depression often indicates a more serious psychological state that can be ongoing and require extensive intervention. In developing a predictive model that can effectively determine the presence of depression or signs of depression, we are effectively aiding universities' efforts to identify

students at risk and intervene earlier before the condition can cause long-term academic and social damage.

One of the goals in this project was to use dimensionality reduction to map the various features onto a 2-D graph in order to visualize the clustering of the three main mental health conditions in the dataset. This was to provide a measure of comorbidity, or how much overlap existed between the conditions. The technique in use is called t-SNE (t-Distributed Stochastic Neighbor Embedding), which allows us to compress the dimensions– demographics, academic factors, PSS-10, GAD-7, and PHQ-9 items– into a 2-D map that showed clusters of observations. The first graph shows the t-SNE visualization for the target class of "Severe Depression," which illustrated similarity measurements between observations (Figure 1). This means that if two survey responders gave the same or similar responses to the stress, anxiety, or depression questions, they would appear closer to each other in the t-SNE plot. Overall, the graph shows a clear gradient progression as depression severity worsens, which informs us that the diagnostic questions effectively separate clusters of students with differing depression levels. A second plot used RGB color channel visualization, which effectively displayed the comorbidity of stress, anxiety, and depression within the dataset (Figure 2). The various mixed colors for each data point on the plot inform us of what we already assumed about the data– there is high comorbidity, where the survey responders suffering from more severe depression levels also suffer from high anxiety or stress levels. These comparisons imply that this dataset has a good structure with which to build machine learning models and from which to learn valuable information about mental health.

We applied some necessary preprocessing steps to ensure the effective training of our models. The dataset at publication was already cleaned of missing or invalid values, so there was

nothing that we needed to impute. We needed to drop several columns from the training dataset. First, the university column was dropped because we did not want to train the model to use the student's university as one of the main predictors. That is, if we intend to use this model for universities and student populations beyond just the top 15 universities in Bangladesh, this feature needs to be excluded. The labels for stress and anxiety were dropped because they were redundant with stress and anxiety values. We needed to drop the depression value column since it directly encoded our target column of depression label. We also needed to drop four out of the nine PHQ questions (PHQ's 1, 2, 6, and 9). These were the most impactful giveaway questions that directly ask the responder if they are feeling depressed or discuss topics like suicidal ideation or lack of pleasure in everyday life. Basically, these questions constitute diagnostic criteria that very easily leaks information about the target class for the given observation. As such, they needed to be dropped in order to make sure that the model was learning realistically and not off of clear-cut cases. These efforts allowed us to prepare a dataset that we could use to build models that both met the goals of the project and were capable of generalizing on new observations.

3.2 Model Training Approach

The first model trained was a logistic regression model, which we used as a baseline for predictive performance metrics. This choice was desired because logistic regression is simply classification using the linear model architecture to convert a weighted sum of the input features into a set of probabilities for each class. This model is therefore less of a black box than the complex XGBoost model that we compared it against. It is also much better for this dataset as opposed to more simple models using Naïve Bayes because we know from the previously listed t-SNE plots that the features are highly correlated. We built and trained this model using a

scikit-learn pipeline with a preprocessor that standardized numerical features and converted categorical features using one-hot encoding. To identify the most effective hyperparameter configuration for the model, we implemented a GridSearchCV that trained a set of candidate models using cross validation, comparing each model's performance, each using a different L2 regularization penalty. We chose the "saga" solver, as it is a fast algorithm in terms of model training time and supports multinomial classification and L2 regularization.

The second and more robust model that we trained was an XGBoost classifier. This was similarly wrapped inside of a scikit-learn pipeline, which allowed us to reuse the same preprocessor as in the logistic regression model from the previous step. The only difference was the usage of label encoding for the target class, since XGBoost requires the target to be represented as numeric class indices. XGBoost is a homogeneous ensemble of decision trees built sequentially, where each tree is grown off of the gradient of the previous one. To do hyperparameter optimization and comparison with this model, we applied a Bayesian search, which uses a probabilistic model to measure how much each hyperparameter influences the model's predictive accuracy. This has allowed us to compare several hyperparameters, each with a wide range of values, within a tight search space that required few evaluations, thereby significantly reducing the training time. Decision trees such as in XGBoost are prone to overfitting, so to ensure the model generalized well, we restricted the range of certain hyperparameters such as the max tree depth or the number of trees. This process yielded an XGBoost model that was capable of capturing the more complex nonlinear relationships that the previous logistic regression model potentially failed to represent.

3.3 Results

Both the logistic regression and XGBoost models appeared to perform relatively well, with an accuracy of about 72% for logistic regression and accuracy of around 73-74% for XGBoost. Several rounds of training resulted in the selection of different configurations of hyperparameters each time, which suggests that stronger performing setups for each model could potentially still be found through further tuning or extra attempts. Despite this variation, the confusion matrices for both models were practically identical (Figure 3). XGBoost on average produced F1 and accuracy scores that were greater than that of logistic regression, but the improvement was only marginal– within a 1% difference in predictive performance. We had expected XGBoost to outperform logistic regression in this project, but the primary reasons for this model's underperformance were the size of the MHP dataset and the presence of strongly dominating PHQ items. Both of these characteristics meant that the dataset surprisingly had fewer nonlinear relationships that could be represented or broken up into several tree structures. Essentially, XGBoost's strengths at modeling these relationships via tree-splitting were not leveraged with this dataset, allowing a simpler classifier to perform almost equally. These results suggest several possible avenues for improving model performance. The first is to simply eliminate all of the PHQ items in the preprocessing step, which would enable the model to learn from more of the factors related to anxiety and stress or from the sociodemographic features. Another option is to explore stacked ensembles where we can combine the strengths of both logistic regression and XGBoost, which might help performance by being able to capture both dominant linear relationships and complex nonlinear relationships.

The results of highest interest are the SHAP explainability scores that were assigned to the features to determine the impact that each individual feature had on the classifier's

predictions. We chose the XGBoost model for these explanation scores because, in contrast to logistic regression which operates using the more basic linear model, XGBoost is a black-box machine learning model with internal decision processes that are more difficult to interpret or trace. As highlighted in the literature, predictions from black-box models stand to benefit substantially from explainability methods like SHAP, since it can provide a "deeper dive into the data for interpreting feature impact" and can more clearly "allocate credit for the model output among its input features" (Wang, et al., 2021, para. 9). Our SHAP summary results, which are organized by the target class of depression severity, provided a comprehensive overview of how much each feature contributes to each of the model's predictions (Figure 4). As expected, the most impactful features are the PHQ items, five of which occupy the top rankings in the beeswarm plot (Figure 5). This makes sense because the PHQ questions have the strongest linear relation with the target class, and this can be observed with those features having long-trailed distributions in the plot. This means that high values for those features increased predicted depression severity, whereas low values suppressed it. The moderately predictive features include the anxiety (GAD) and stress (PSS) items, which is consistent with what we can expect with the comorbidity of depression with both of these secondary mental conditions. Interestingly, the primary academic indicator of cumulative grade point average (CGPA) and the gender feature also appear in the SHAP listing, meaning they have similarly moderate importance. Mirroring the aforementioned study from Nanjing, the SHAP analysis allows us to distinguish the most important features, thereby giving us several actionable insights into academic mental health.

4 Conclusions

This study set out to use machine learning to model depression risk factors among university students using the MHP dataset and leveraged SHAP AI explainability to obtain a measure of which risk factors were the most salient. Our results demonstrated that logistic regression and XGBoost performed very similarly, but SHAP provided deeper interpretability to the XGBoost model, following its use case as suggested in the literature. This study supports the broader academic movement towards providing more explainability to AI and machine learning. By doing so, we have arrived at very meaningful insights, which healthcare professionals and universities beyond those specific to this study can use to detect and prevent academic mental illness. Specifically, the higher priority items identified by the SHAP analysis should prompt universities to target PHQ diagnostic criteria in interventions to identify and aid students at risk for severe depression. This includes but is not limited to providing lifestyle coaching related to sleep hygiene or quiet study hours, setting up exercise, nutrition, and wellness workshops, or offering stress and anxiety counseling events. This also may involve identifying academic struggle through monitoring students' CGPA or offering more early-semester support or safety nets for suffering students. Future improvements on this project includes expanding the dataset to incorporate more universities from around the world and exploring further models that may have higher accuracy or generalizability. Ultimately, this project was successful in blending the initiatives of data science and mental health research and contributing to broader efforts to create actionable AI systems in healthcare.

References

Auerbach, R. P., Alonso, J., Axinn, W. G., Cuijpers, P., Ebert, D. D., Green, J. G., … Bruffaerts, R. (2016, August 3). *Mental disorders among college students in the World Health Organization World Mental Health Surveys: Psychological Medicine*. Cambridge Core. https://www.cambridge.org/core/journals/psychological-medicine/article/mental-disorders-among-college-students-in-the-world-health-organization-world-mental-health-surveys/34942DEAFC35899349114B73E84FB080

Ibrahim, A., Kelly, S., Adams, C., & Glazebrook, C. (2023, March). *A systematic review of studies of depression prevalence in university students*. ScienceDirect. https://www.sciencedirect.com/science/article/abs/pii/S0022395612003573?via%3Dihub

Iyortsuun, N. K., Kim, S.-H., Jhon, M., Yang, H.-J., & Pant, S. (2023, January 17). A review of machine learning and deep learning approaches on mental health diagnosis. PMC PubMed Central. https://pmc.ncbi.nlm.nih.gov/articles/PMC9914523/

Syeed, M., Rahman, A., Akter, L., & Fatema, K. (2024, May). *A Comprehensive Standardized Dataset on Mental Health Problems (MHPs) of University Students*. A comprehensive standardized dataset on Mental Health Problems (MHPs) of University Students. https://www.researchgate.net/publication/380638187_A_comprehensive_standardized_dataset_on_Mental_Health_Problems_MHPs_of_University_Students

Wang, C., Feng, L., & Qi, Y. (2021, November). *Explainable deep learning predictions for illness risk of mental disorders in Nanjing, China*. ScienceDirect. https://www.sciencedirect.com/science/article/abs/pii/S0013935121010343

Appendix

Figure 1

t-SNE Similarity Clustering color-coded to target class

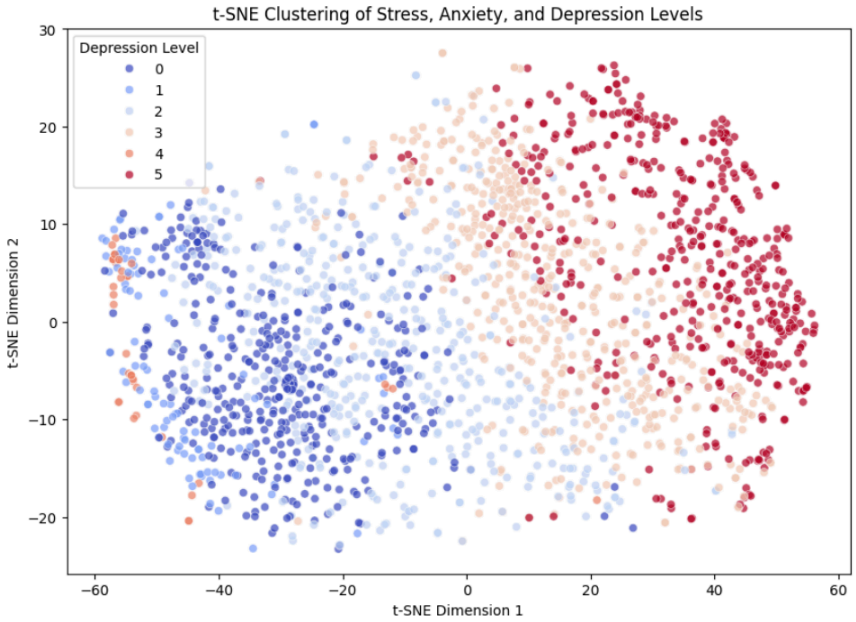
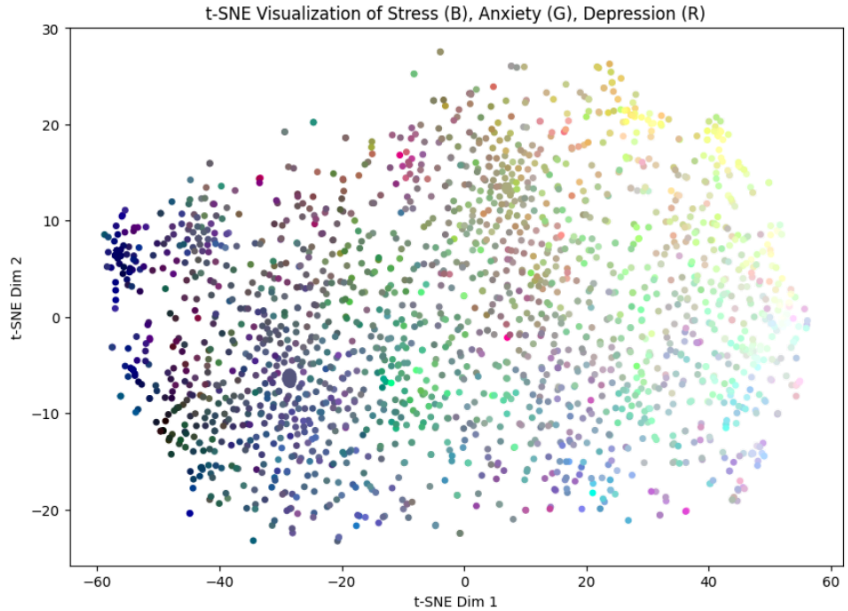


Figure 2

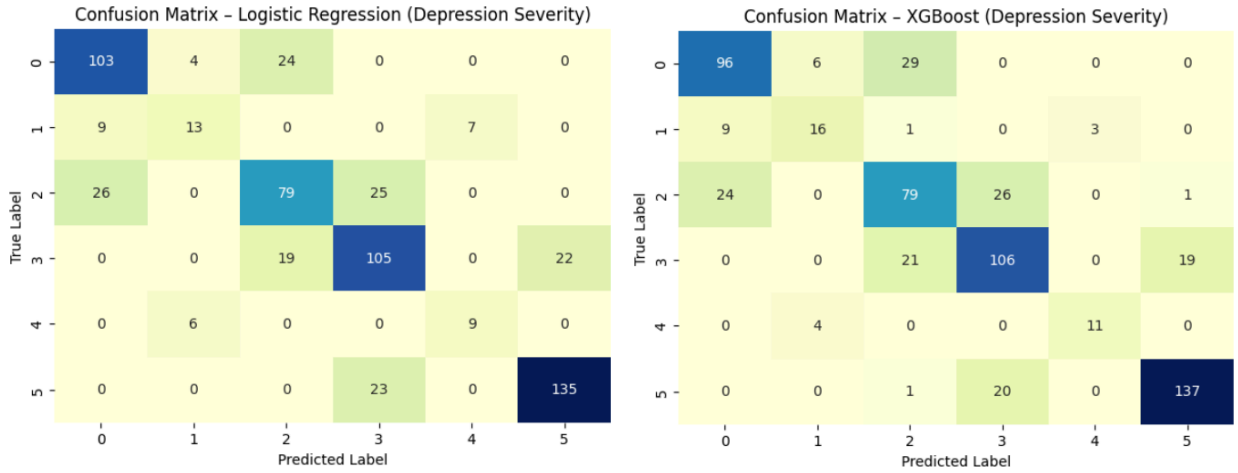
t-SNE Similarity Clustering color-coded according to comorbidity levels of anxiety, stress, and depression



Note: Red channel corresponds to depression. Green channel corresponds to anxiety. Blue channel corresponds to stress.

Figure 3

Confusion matrices for logistic regression and XGBoost models.



Note: Scores are as follows for the two models:

Logistic Regression (on average):

Accuracy.....: 72.9064

Precision.....: 72.7160

Recall.....: 72.9064

F1 score.....:0.7274

XGBoost (on average):

Accuracy.....: 73.0706

Precision.....: 73.1044

Recall.....: 73.0706

F1 score.....:0.7307

Figure 4

Multi-class SHAP bar plot showing overall feature importance

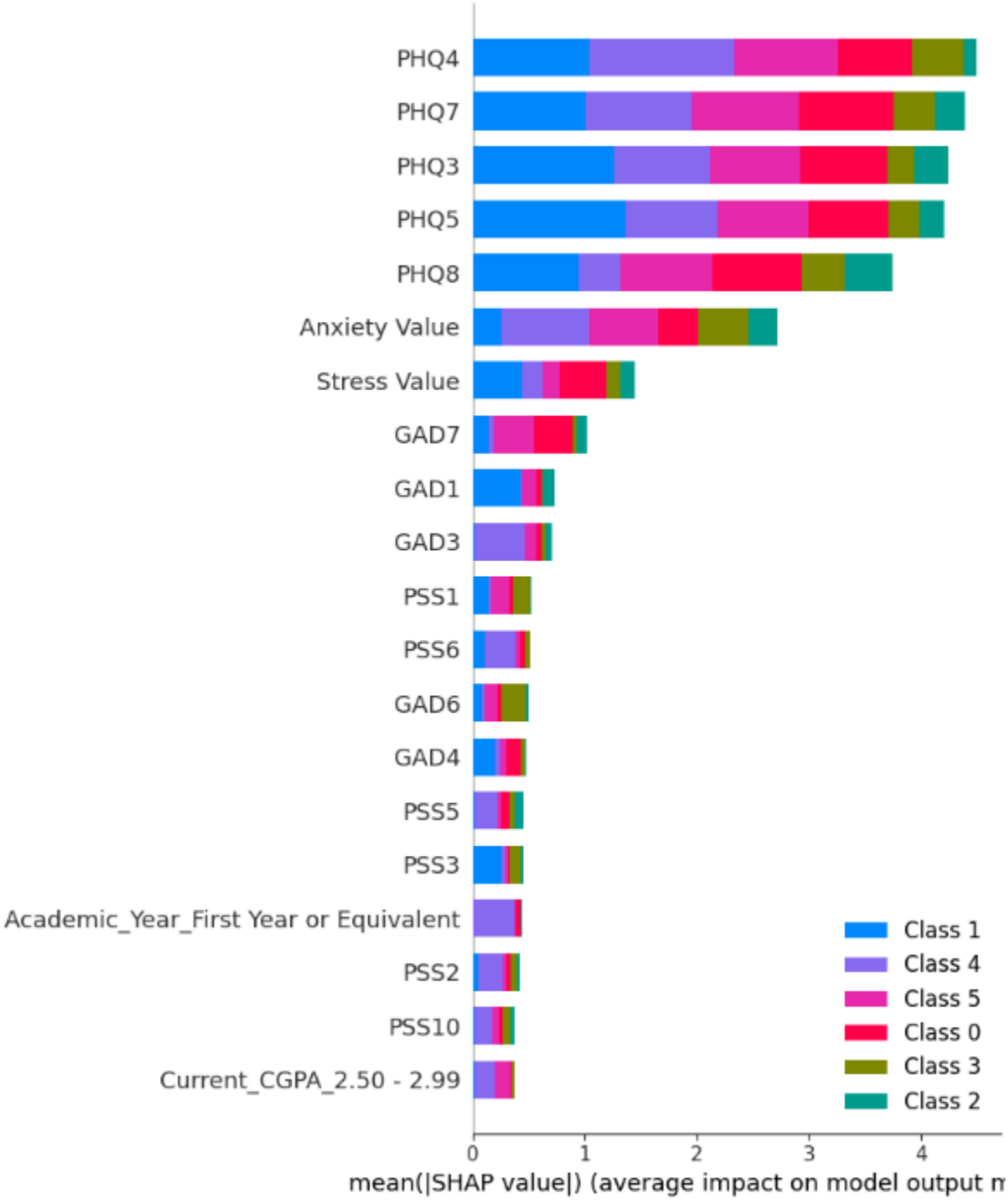


Figure 5

Beeswarm plot showing distribution of questions responses corresponding to prediction impact