Justin Lee (jgh2xh) and Bryan Christ (brc4cb)

Predicting Anxiety and Depression Among K-12 Teachers in Latin America During the COVID-19 Pandemic

Abstract

In this project we aim to examine how well different machine learning models can predict anxiety and depression amongst K-12 teachers in Latin America during the COVID-19 pandemic. Further, we look to compare these results to logistic regression, a more traditional approach. We find that logistic regression does best for predicting anxiety, and random forest is most effective for depression. Interestingly, we find that predicting anxiety is a more difficult task than predicting depression in our sample. In both cases, however, no one model performed unilaterally better than the others across the metrics we considered. From this we conclude that it is worthwhile to have these multiple models to select the model with the best performance in the most relevant metric for the required task.

Data and Methods

Dataset and Preprocessing

The dataset was gathered via surveys of K-12 teachers in Latin America during the COVID-19 pandemic. The dataset consists of 2,004 rows and 50 predictors. After dropping all collinear variables, the resulting dataset contains 45 predictors, which can be broadly subdivided into six different categories: demographics, personal health, educational background, teaching background, COVID-related variables, and teaching during COVID variables. Most variables considered are either binary (e.g., rural vs urban) or ordinal (e.g., COVID fear measured on a Likert scale), but a few are discrete (e.g., age).

The outcome variables of depression and anxiety are initially encoded as ordinal variables, but we re-encoded these as binary variables using 1 to represent clinically relevant scores (3 and above on a scale of 0-6) for depression and anxiety, and 0 for non-clinically relevant scores. We chose to binarize the outcome variables because it allows us to translate the task into a classification objective, for which we can easily compare logistic regression to more advanced machine learning classification algorithms. We used a 85-15 train-test split in order to keep the training set as large as possible without reducing the usefulness of the test set. The resulting training data was heavily unbalanced with only 37% of the rows containing anxiety, and an even lower 14% for depression. In order to remedy this, we upscaled the training data with respect to the target variable (anxiety or depression, depending on the classification task) by sampling with replacement enough times as necessary to balance the outcome

label. Crucially, we also include depression as a predictor for anxiety and vice versa because the two are greatly correlated.

Methods

We compared the results from five different models: our baseline logistic regression, random forest (RF), multilayer perceptron (MLP), support vector machine (SVM), and gradient boosted tree (GBT). The evaluation metrics we used are: accuracy, precision, recall, F1, and area under the receiver operating characteristic curve (AUROC). In order to retain comparability of the results, we opted to use the same pipeline for each of the models. The only preprocessing we apply is to standardize the variables to have mean 0 and variance 1.[1] This data is then passed into the model for training. We do a hyperparameter grid search, and validate the results using four-fold cross-validation. The hyperparameter grids are listed in Table 3. The final performance metrics are calculated using the best model on the held out test set.

Results

Anxiety Models

As shown in Table 1, none of the five models considered had a high level of accuracy when predicting anxiety in the holdout test dataset. The most accurate model was logistic regression, with an accuracy of 76.11%. With the exception of GBT, which had an accuracy 7 percentage points below logistic regression, the other models considered had similar accuracy to logistic regression. Across the other four evaluation metrics, the gap in performance between the highest and second highest performing models is relatively small (3 percentage points for recall but within 1 percentage point for precision, AUROC, and F1), suggesting no model has a clear advantage in any of these metrics.

Based on its performance across each of the five evaluation metrics, we conclude that logistic regression is the most effective model for predicting anxiety in this sample. Logistic regression not only has the highest overall accuracy, but also best balances the tradeoffs between false positives and false negatives, as reflected in the model having the highest F1 and AUROC scores at 80.42% and .7480, respectively. Considering the balance between false positives and false negatives is important when predicting psychological outcomes because both under and over predicting the outcome could have negative consequences for patients. With under prediction, we fail to provide supports and medication to those experiencing adverse psychological outcomes, whereas with over prediction, we might provide medication that alters the biochemistry of healthy individuals.

[1] We skipped this step for the tree-based methods RF and GBT.

Based on our results, anxiety is a difficult construct to predict in this sample regardless of the model considered. We argue that it does not appear to be necessary to use more advanced machine learning algorithms for predicting anxiety in this sample, as the regression-based approach performed the best overall.

Depression Models

As shown in Table 2, unlike with anxiety, RF had the highest level of accuracy for predicting depression in the holdout test dataset by a large margin (6 percentage points) with an accuracy of 86.62%. RF also had the highest recall and F1 scores, leading us to conclude it is the most effective model overall because it best balances true positives and false negatives (recall) as well as the overall balance between false positives and false negatives (F1). As mentioned above, it is important to consider both under and over prediction when predicting mental health outcomes, and RF best addresses this tradeoff. One drawback of the RF model is that it has a fairly low AUROC (.7054), suggesting it will not perform as well at different prediction thresholds than the traditional binary classification cutoff of having a predicted probability greater than .5. However, this is a relatively minor concern given the model performs well at the .5 prediction threshold, so it is likely not necessary to adjust this cutoff.

One interesting finding is that while the other four models considered struggled with overall accuracy, each performed well in one or more of the other metrics considered. For example, SVM had the lowest accuracy, but the second highest precision (98.45%) and AUROC (.8132). Thus, while we argue that RF is the best model for predicting depression in this dataset, it is important to consider other models when predicting psychological outcomes as they each come with their own unique tradeoffs. Considering our results as a whole, it is clear depression is more easily predicted using the combination of variables in this dataset than anxiety.

Hyperparameter Tuning and Sensitivity for Best Models

As shown in Table 4, according to four-fold cross-validation via grid search, our best performing logistic regression model for predicting anxiety had an elastic net parameter of 1.00 and a regularization parameter of .01. Sensitivity analysis revealed that this model was not very sensitive to different hyperparameter combinations, as all models had a cross-validated accuracy between 80-83%. As shown in Table 4, our best performing RF model for predicting depression had a max depth of 15 and 45 trees. Unlike with logistic regression, the RF model was very sensitive to different hyperparameter values, with a minimum cross-validated accuracy of 79.02% (max depth 1, 10 trees) and maximum cross-validated accuracy of 99.87%.

Conclusion and Future Research

Based on our results, we argue it is possible to predict psychological outcomes using machine learning models with a good level of precision, as we were able to predict depression with 86.62% accuracy. Our results also make it clear that it is important to consider multiple models and metrics, as there was no clear winning model across all metrics for predicting anxiety and depression. Considering multiple models and metrics will also help clinicians and psychologists best balance the tradeoff between over and under prediction depending on the context.

We also argue that future research should include additional variables that might lend more explanatory power to and increase the predictive accuracy of machine learning algorithms. Indeed, perhaps our models would have been more accurate had the dataset included additional predictors that might be closely correlated with anxiety and depression such as geographic location, social support, a more nuanced measurement of socioeconomic status (our dataset only included a categorical measurement of socioeconomic status), and more student/school characteristics (i.e., student income level, measures of school funding, and student outcomes such as test scores and disciplinary metrics). Given the fact that logistic regression performed the best for predicting anxiety, future research should include both advanced machine learning models and traditional regression-based approaches.

Appendix A

List of Variables

Demographic Information (4)

- Age
- Socioeconomic Status
- Sex
- Partnered

Educational Background (7)

- High School Diploma
- Bachelor's Degree
- Master's Degree
- PhD
- Postdoctoral Fellowship
- Non-university Studies
- Education Degree

Personal Health (5)

- Pre-Pandemic Chronic Illness
- Pre-Pandemic Mental Illness
- Pre-Pandemic Neurological Disability
- Anxiety/Depression
- Overall Health

Teaching Background (9)

- 0-5 Years
- 6-11 Years
- 12-16 Years
- 17-18 Years
- Rural vs Urban
- Secondary vs Primary
- Public vs Private
- Years as Teacher
- Employed vs Unemployed

COVID Related (4)

- Past COVID Positive
- COVID Vaccinated
- COVID Fear
- Sufficient COVID Measures

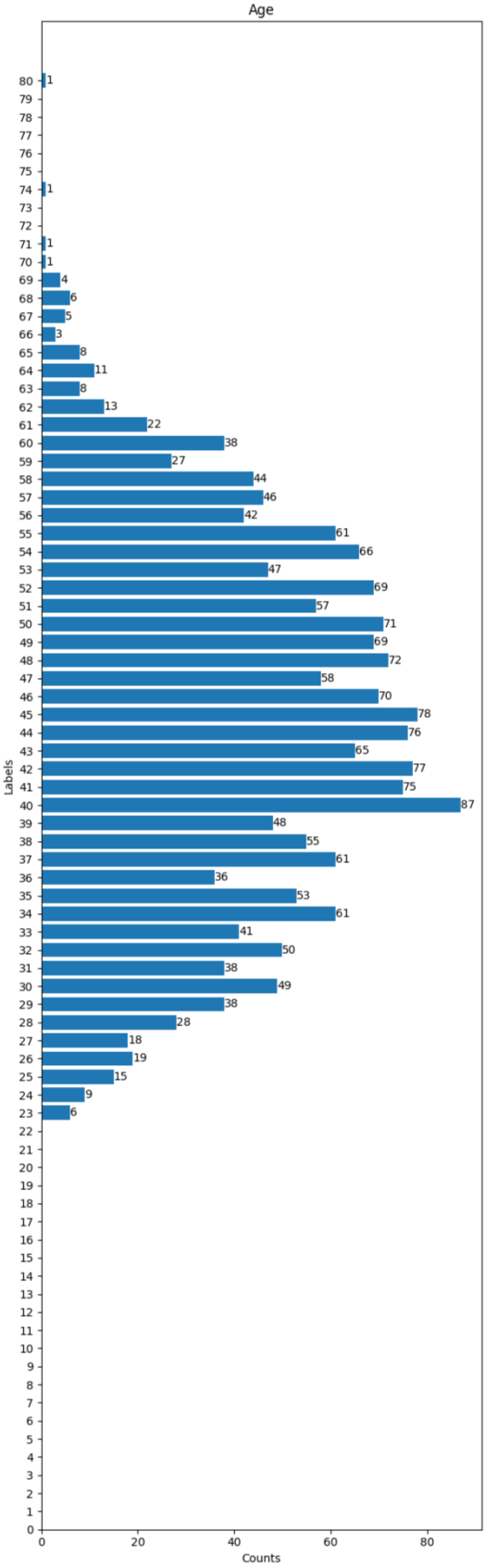
Teaching During COVID (16)

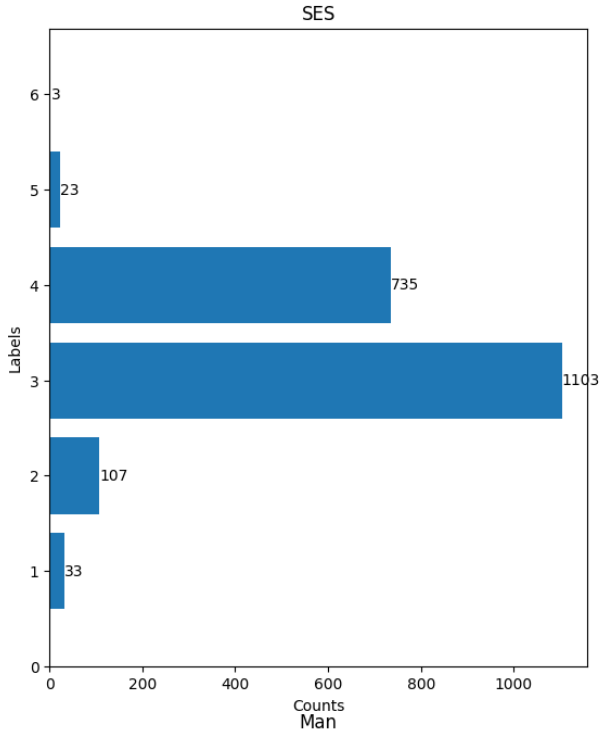
- Relationship Improvement with Students
- Workload Now vs Pre-COVID
- Resource Satisfaction
- Months of Online Teaching
- Student Educational Problems
- Student Behavioral Problems
- Student Emotional Problems
- Student Social Problems
- Student Family Problems
- Difficulty of Online Teaching
- Instructional Adjustment
- Benefits of Online Teaching
- Total Online Teaching
- Maslach Burnout: Emotional Exhaustion
- Maslach Burnout: Depersonalization
- Maslach Burnout: Fulfillment

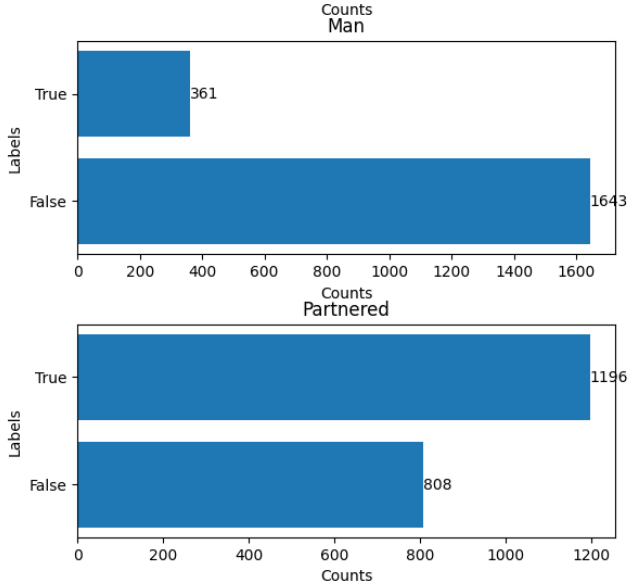
Appendix B

Plots of Select Variables

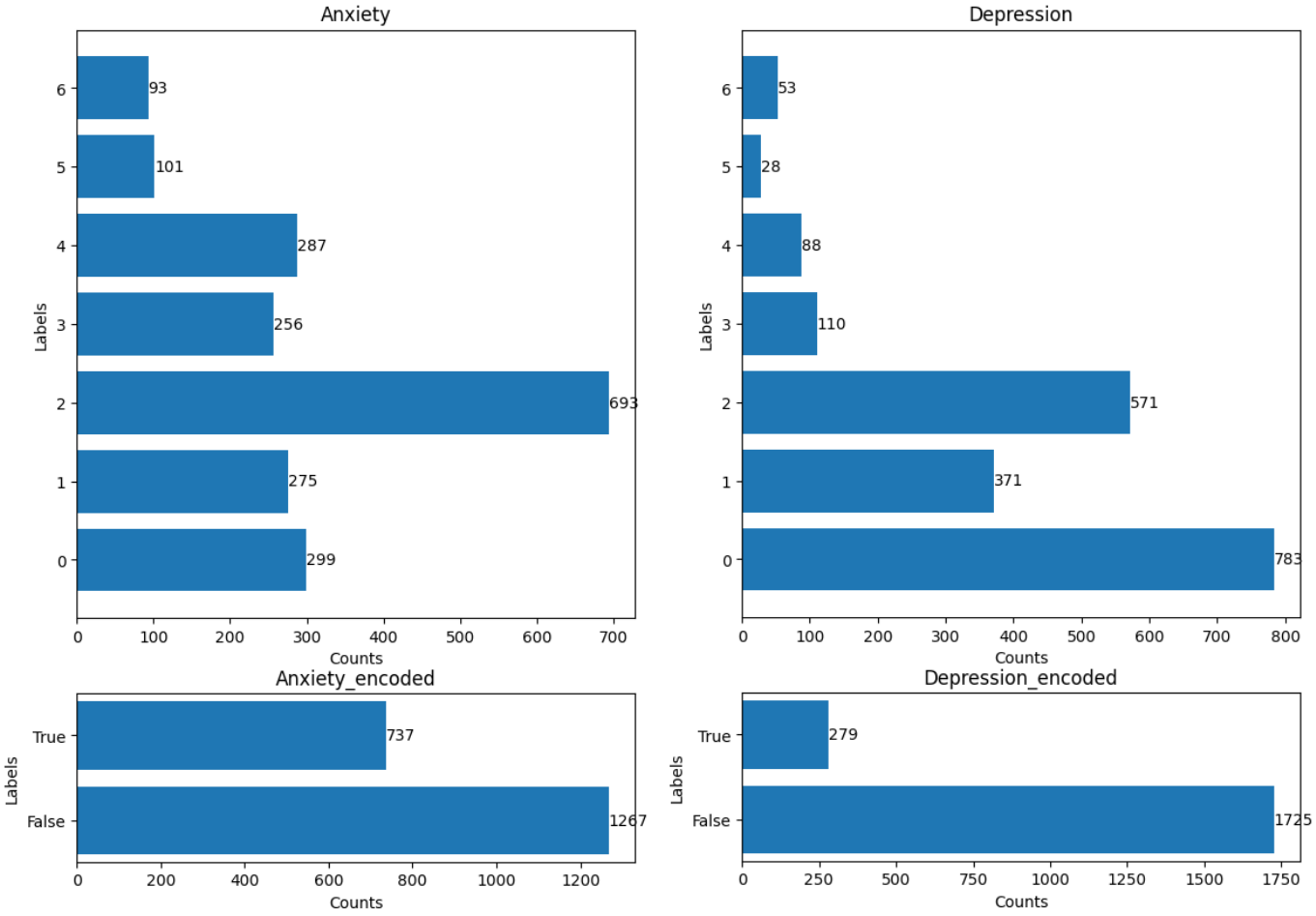
Demographic Information



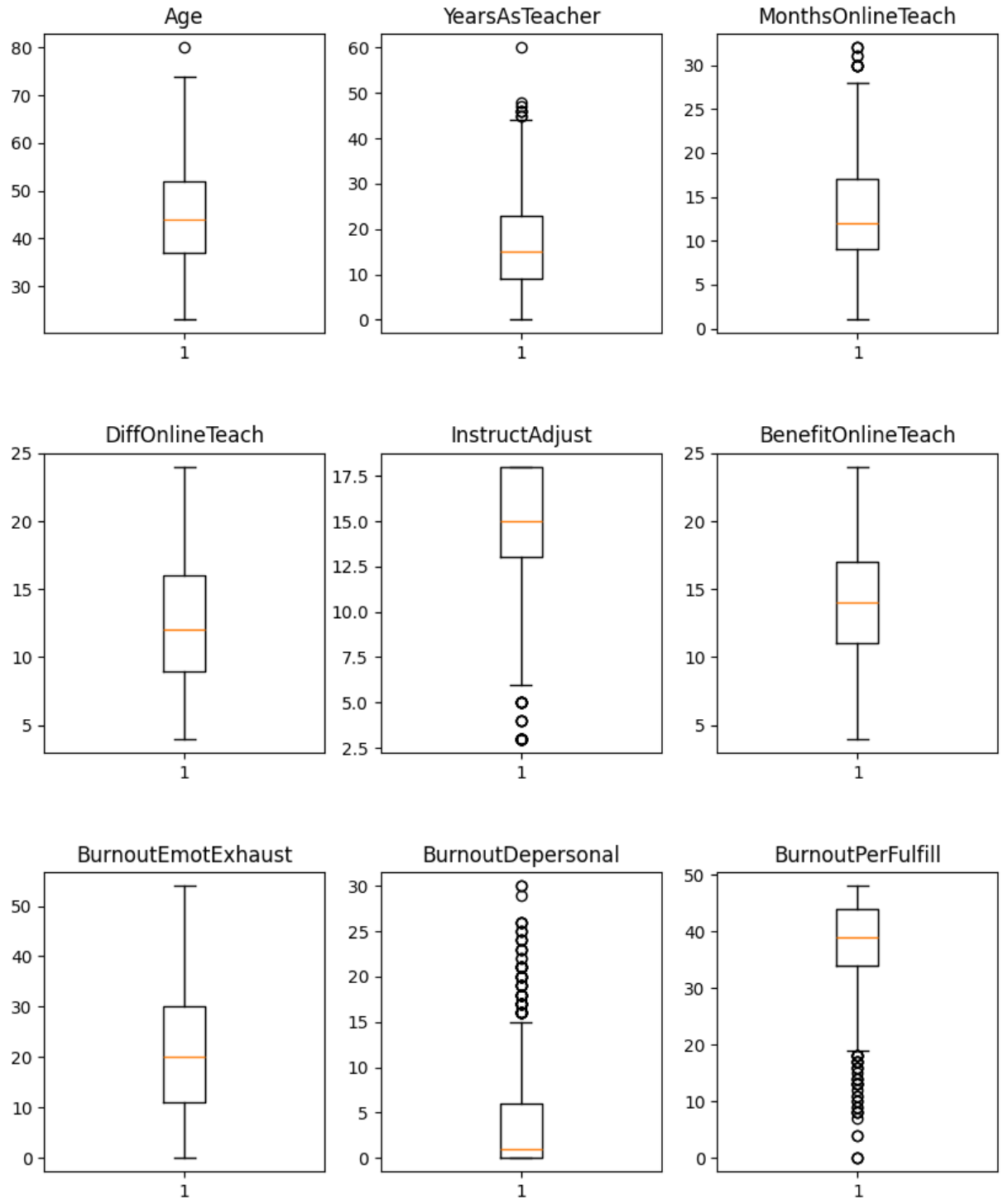




Target Variables Before and After Encoding



Boxplots of Select Discrete Variables



Appendix C

Table 1

Anxiety Results

	Accuracy	Precision	Recall	F1	AUROC
Logistic Regression	**0.7611**	0.7979	0.8105	**0.8041**	**0.7480**
RF	0.7357	0.7488	**0.8474**	0.7951	0.7059
MLP	0.7452	**0.8090**	0.7579	0.7826	0.7419
SVM	0.7548	0.7897	0.8105	0.8000	0.7399
GBT	0.6911	0.7360	0.7632	0.7494	0.6719

Note. This table displays the results of each of the five models considered across five metrics of accuracy for predicting anxiety in the holdout test dataset. Bolded numbers represent the highest score for each metric, while the bolded model represents the best overall model for predicting anxiety.

Appendix D

Table 2

Depression Results

	Accuracy	Precision	Recall	F1	AUROC
Logistic Regression	0.7484	0.9755	0.7289	0.8343	0.8035
RF	**0.8662**	0.9231	**0.9231**	**0.9231**	0.7054
MLP	0.7452	**0.9898**	0.7143	0.8298	**0.8328**
SVM	0.7293	0.9845	0.6996	0.8180	0.8132
GBT	0.8057	0.9206	0.8498	0.8838	0.6810

Note. This table displays the results of each of the five models considered across five metrics of accuracy for predicting depression in the holdout test dataset. Bolded numbers represent the highest score for each metric, while the bolded model represents the best overall model for predicting depression.

Appendix E

Table 3

Hyperparameter Grids for Each Model

	Hyperparameter Grid
Logistic Regression	Elastic Net Parameter: [0, 0.2, 0.4, 0.6, 0.8, 1] Regularization Parameter: [0.1, 0.01, 0.001]
RF	Max Depth: [1, 5, 10, 15] Number of Trees: [10, 15, 20, 25, 30, 35, 40, 45]
MLP	Layers and Neurons: [45, k, k, 2] for k in [50, 75, 100] Step Size: [0.001, 0.005, 0.01, 0.05]
SVM	Max Iterations: [10, 20, 30, 40, 50] Regularization Parameter: [0.001, 0.01, 0.1]
GBT	Max Depth: [5, 10, 20, 30] Max Bins: [8, 16, 32, 64]

Appendix F

Table 4

Best Hyperparameter Values for Each Model

	Anxiety	Depression
Logistic Regression	Elastic Net Parameter: 1.00 Regularization Parameter: 0.01	Elastic Net Parameter: 0.00 Regularization Parameter: 0.01
RF	Max Depth: 15 Number of Trees: 40	Max Depth: 15 Number of Trees: 45
MLP	Layers and Neurons: [45, 100, 100, 2] Step Size: 0.05	Layers and Neurons: [45, 100, 100, 2] Step Size: 0.05
SVM	Max Iterations: 10 Regularization Parameter: 0.01	Max Iterations: 30 Regularization Parameter: 0.001
GBT	Max Depth: 10 Max Bins: 32	Max Depth: 10 Max Bins: 64

Note. This table displays the best hyperparameter values for each of the five models considered for both anxiety and depression. The hyperparameter values were determined using 4-fold cross validation via grid search using the values defined in Table 3.