Dylan Hill and Wissal Khlouf
Javier Rasero
DS 4021
10 December 2025

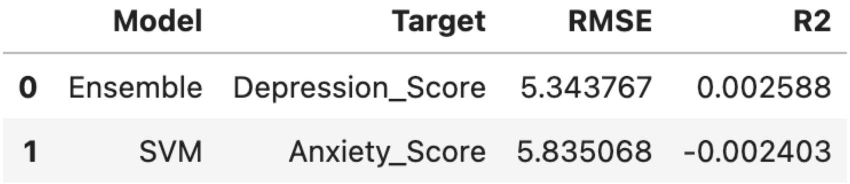
Final Report

Anxiety and depression remain among the most widespread and costly mental health challenges worldwide, despite the extensive research conducted on these disorders each year. Their impact is profound: more than one billion people currently live with psychological health issues, and depression and anxiety alone are estimated to cost the global economy over US$1 trillion annually, with suicide claiming over 700,000 lives in 2021. While biological and genetic information is known to play a substantial role—approximately 40–50% of depression risk is attributed to genetics—such data are often inaccessible, particularly for low-income populations who are 1.5 to 3 times more likely to experience mental illness and face a self-reinforcing cycle between poverty and psychological distress. Because the individuals most at risk are also those least able to obtain advanced diagnostic resources, developing predictive tools that do not rely on biological data is a pressing challenge. In this project, we investigate the following research question: How well can anxiety and depression be predicted using only accessible, non-biological features? By evaluating machine-learning models on these alternative predictors, we aim to better understand whether reliable forecasting is possible without specialized medical information.

The dataset we used for this project came from Kaggle, it is called ‘Anxiety and Depression Mental Health Factors’. There are 1,200 rows of data from 1,200 participants, as well as 21 columns in total, including an anxiety score and a depression score, both of which we were trying to predict with the remaining features. In our descriptive analysis notebook, one of the first things we did - before producing visualizations and summary statistics describing the relationships between columns in the data - was check to see if any of the columns had missing values. We quickly observed that both the ‘Medication_Use’ and ‘Substance_Use’ columns had lots of missing values - in fact, the majority of the rows in those columns were missing. We decided that we had two options, which were to either remove all the rows in the entire dataset that were missing a value thanks to one (or both) of those columns, or simply remove those columns from the dataset. We tested each option and saw that by getting rid of the rows with the missing values, we would be dwindling the dataset down to just 140 rows. We decided that this was not the way to go as we figured this would not be enough data to do any meaningful analysis with, so we got rid of the two columns and kept on moving forward.

To summarize our predictive modeling process, each model was trained and optimized twice: once using depression scores as the target variable and once using anxiety scores. For the ensemble approach, we implemented a Random Forest model, embedding all preprocessing within the cross-validation pipeline as required. This preprocessing step primarily involved encoding categorical variables (and would have addressed missing values had any remained), after which we performed hyperparameter tuning through GridSearchCV to optimize the number of trees, maximum tree depth, and number of features considered at each split. The best-performing hyperparameter combination—selected based on mean squared error—was then refitted on the training data, and we evaluated the final model using 5-fold cross-validation, reporting RMSE, MAE, and R^2. For the neural network model implemented in PyTorch, the workflow followed a similar structure but required additional preprocessing: categorical variables were encoded and numeric predictors were standardized, as scaling is essential for neural network optimization. We defined a custom network architecture using nn.Module and created a fold-level training function using MSE loss and the Adam optimizer. Hyperparameter tuning focused on learning rate, hidden layer size, and number of epochs, with performance compared using average RMSE across folds. After identifying the best hyperparameters, we trained the final model and evaluated it with outer 5-fold cross-validation, reporting RMSE and R^2 for the optimized network. For our penalized linear model, we implemented Ridge regression separately for predicting depression and anxiety. We embedded preprocessing inside the same cross-validation pipeline used for the other models, including one-hot encoding for categorical variables and standard scaling for numeric features to prevent data leakage between folds. Hyperparameter tuning was performed using GridSearchCV over a range of regularization strengths (alpha values), with five-fold cross-validation and negative mean squared error as the optimization metric. Model performance was summarized using RMSE, MAE and R^2. For the Support Vector Machine model, we similarly trained two separate SVR models, one for depression and one for anxiety, again embedding preprocessing within the pipeline and tuning hyperparameters including penalty parameter C, kernel type, and gamma. Cross-validation was performed using five folds with the same evaluation metrics, allowing us to directly compare performance across both targets.

After selecting the best performing models from the cross-validation phase, we refit those models on the full training data and evaluate them on the test set. For predicting depression, the Random Forest ensemble achieved the lowest RMSE (≈5.34) and a near-zero R^2 value, indicating that it performed slightly better than a naive baseline but still explained virtually none of the variance. For anxiety, the Support Vector Machine was the best performing model, but similarly showed low predictive power (RMSE ≈5.83 and slightly negative R^2). The table summarizes the test-set performance for each best-performing model for both targets. Overall, these results indicate that the models were roughly consistent in their magnitude of error, but that none could meaningfully predict individual scores.



	Model	Target	RMSE	R2
0	Ensemble	Depression_Score	5.343767	0.002588
1	SVM	Anxiety_Score	5.835068	-0.002403

Across models, the performance was uniformly low. This may indicate that non-biological predictors have a relatively low predictive signal to contribute to the modeling of complex psychological outcomes like anxiety and depression. The ensemble model may have performed slightly better for the depression outcome because tree-based methods are more suited to capture non-linear interactions among features. The highest performance for anxiety was observed with SVM, which may be a reflection of SVM's ability to fit flexible decision boundaries when hyperparameter-tuned with non-linear kernels. On the other hand, it is worth noting that, overall, R^2 values were near-zero or negative, which indicates that the models were practically unable to learn any relationship beyond predicting values around the mean of the dataset. A significant shortcoming of this project is that the data itself and the specific predictors available in the dataset are major limiting factors. Self-reported survey variables limited demographic information, and exclusion of biologically meaningful features from the data likely constrained the information available for prediction. The results of this analysis further support the position that while accessible features are important, biological and clinical data seem to be necessary to generate clinically useful performance for mental health.