The Susceptibility of Anxiety, Depression, and ADHA in Adolescents

Shaili Gunda, Bill Lu, Mihir Padsumbiya, Meet Trada, Eunice Choe

The University of Texas at Dallas, Naveen Jindal School of Management

HMGT 6334/ MIS 6305 Healthcare Analytics

Professor Mehmet U.S Ayvaci

November 29, 2023

Executive Summary:

For this project, we wanted to create a predictive model pertaining to mental health disorders in US minors under 18. Mental health illnesses are on the rise, and we wanted to focus on the demographic where intervention is most impactful. To facilitate this goal, we decided to limit our scope to the three most common mental illnesses for children and teens: ADHD, Anxiety, and Depression. With this predictive model, schools and districts will be able to input student data and determine which students are at higher risk of developing these three disorders. We hope this model can be used to facilitate school-based prevention programs and help education officials better allocate mental health resources.

We first started this project by obtaining Client Level Data from SAMHSA (Substance Abuse and Mental Health Service Administration) regarding demographics, medical characteristics, and nine different mental health conditions. SAMHSA obtained this data by compiling patient data provided by the Mental Health Authorities of each state and the patients they treated. The dataset initially had more than 6 million records and 40 variables, so we filtered out all records above 18 years of age and kept only 13 key variables. There were still many null values in the dataset, so using a combination of domain knowledge and rule-based techniques, we imputed or omitted most data points and created dummy variables for a few variables where missingness provides predictive power. Finally, we found factors that had a significant impact on the mental health of minors and attached supplemental data to the original dataset. All the supplemental data came from the result of CDC questionnaires to school districts, and we kept 52 out of the 400+ questionnaire variables.

For model training, an 80 – 20 split was used for 4-way cross validations. A multitude of trainings methods were used including logistic regression, decision trees, support vector machines, random forest, gradient boosting, extreme gradient boosting, and cat boosting. We initially wanted to create a multiclass classification model for the three disorders but decided to implement three distinct models for each illness due to concerns of accuracy. To measure model performance, we decided to use precision and recall as our performance metrics as the additional cost of each screening is less relevant. In the end, random forests were the best performer for all three disorders with AUCs of 0.72 for ADHD, 0.74 for Anxiety, and 0.80 for depression.

Project Background:

Our project tackled a topic that required greater awareness and advocacy. Healthcare plays a considerable role in the U.S. economy, but Mental health is overlooked. The project's main objective was to help facilitate school-based mental health prevention programs. Only 1 out of 3 schools provide outreach services to students in the U.S. Implementation of these programs would improve outcomes through early detection and treatment, allow for greater access to care, and help education offices reallocate resources. SAMHSA, The Substance Abuse and Mental Health Service Administration states a mental illness as "disorders ranging from mild to severe, that affect a person's thinking, mood, and or behavior."

According to the CDC, the most diagnosed mental disorders in children are ADHD, with about 6 million diagnoses; Anxiety, with 5.8 million; Behavioral Problems, with 5.5 million; and Depression, with 2.7 million. These conditions often occur together with 3 in 4 children who have Depression also have anxiety. Next, our team needed to determine why these statistics were vital. The effects of mental disorders have a lasting impact throughout one's life. A study conducted by NIH National Institutes of Health revealed that mental health issues during childhood are associated with detrimental developmental outcomes in young adulthood. A child's opportunities for health and a socially successful life can be negatively altered. Diagnosed children often display impaired mental health, lower life satisfaction, and poorer health-related quality of life as an adult.

Our data was retrieved from SAMHSA, The Substance Abuse and Mental Health Service Administration, an agency within the United States under the HHS, The Department of Health, and Human Services. SAMHSA was established by Congress in 1992 to create great awareness of substance use and mental disorders through leadership, support programs, and resources. The dataset was compiled on Client Level Data (MH-CLD) from state mental health authorities relating to demographics and mental health characteristics. MH-CLD focuses on individual clients, while SAMHSA integrates another dataset based on treatment episodes (MH-TED). The original dataset contained 6 million entries with 40 variables.

Data Description

Data Cleaning/Preprocessing:

The original dataset contained 6.9 million records and 40 categorical variables. For the scope of this project, it was crucial to tackle two main challenges: identifying important variables to maximize the information used for an efficient model and implementing proper techniques to manage a high number of nulls in some critical variables. The records consisted of individuals from all age groups. However, since the project's focus is solely on minors (Under the age of 18), the database was filtered to include only records from three age groups – 0-11, 11-14, and 14-17 years, coded as 1, 2, and 3 respectively in the original dataset. This refinement reduced the number of observations to 1.8 million.

To filter out unimportant variables, a blend of domain knowledge and data understanding was employed. Variables deemed to have negligible to no impact on the mental health of minors were excluded. These are:

1. "YEAR"
2. "VETERAN"
3. “MARRIAGE STATUS
4. “DIVISION”
5. “REGION”
6. “EMPLOY”
7. “DETNLF”

Furthermore, the original dataset records over 9 different mental health disorders. For the scope of this project, 3 of those mental health disorders were chosen as they form many Mental health reasons in minors. These are: ADHD, Depression, and Anxiety. Hence the rest were dropped from the dataset:

1. "TRAUSTREFLG”
2. "CONDUCTFLG"
3. "DELIRDEMFLG"
4. "BIPOLARFLG"
5. "ODDFLG"
6. "PDDFLG"
7. "PERSONFLG"
8. "SCHIZOFLG"
9. "ALCSUBFLG"
10. "OTHERDISFLG

After filtering for important variables according to domain knowledge, the number of important predictors from SAMHSA dataset were:

1. AGE
2. EDUCATION
3. ETHNICITY
4. RACE
5. GENDER

If the client has records of receiving any service from the following:

6. SPHSERVICE: State Psychiatric Hospital Services
7. CMPSERVICE: Community Mental Health Centers
8. OPISERVICE: Other psychiatric inpatient Services
9. RTCSERVICE: Residential Treatment Center
10. IJSSERVICE: Institutes under the justice system
11. SUB: Substance use diagnosis
12. SAP: Substance use problem
13. SMISED: Serious Mental Illness or Serious Emotional Disturbance

Each of the above variables is a categorical variable. Since there is no inherited ordering in any of these variables, each of the variables with more than 2 categories were further converted to binary format using leave-one out One-hot encoding. This resulted in 44 variables.

The original dataset as is contained a considerable number of nulls for some of the most important variables:

EDUC	116570
ETHNIC	228892
RACE	237329
GENDER	5352
SMISED	71173
SAP	258034
LIVARAG	85441

A combination of Rule-based + Domain knowledge approach has been used to impute for Null values for variables where omitting is not an option due to significant loss of observations. However, for variables

like Gender since the number of null records is significantly less as compared to the total size of the dataset, the null records were omitted. For imputation, significantly correlated variables like age and education were used to impute for the latter. For example, it was observed that individuals in the age group 14-17, most had 9-11 years of education. Hence, individuals in the age group 14-17 will NULL education, was imputed using 9-11 years of education. For variables like SMISET and SAP, missing values themselves provide meaning to the model according to the data dictionary. Hence, since it is MNAR (Missing not at random), it was included in the model as a separate dummy.

Supplemental Data:

A number of factors were identified to have a significant impact on Mental Health in Minors using https://www.cdc.gov/childrensmentalhealth/basics.html. Some of the most key factors:

1. Divorced parents
2. School Environment: Violence, Bullying, Gang activity
3. Academic Stress
4. Peer pressure: Substance abuse
5. Poverty
6. Rebelliousness
7. Parental unemployment
8. Parental Drug abuse
9. Sexual abuse

To account for these factors, supplemental data was used to replace state names with their respective scores for a variable using district level aggregated data. CDC organizes questionnaires for school districts to address an exhaustive list of factors. The questionnaires were thoroughly analyzed to filter out 52 variables from over 400+ variables, that could have a direct impact on the Mental Health of a student. A sample of variables chosen are:

1. Divorce Rate
2. Prohibition of Gang Activity
3. Number of Counselors available on school campus
4. Support groups for addressing Mental Health Disorders
5. Resources for Mental Health support staff
6. Prohibition of Cyber bullying
7. Action Plan for suicide risk.

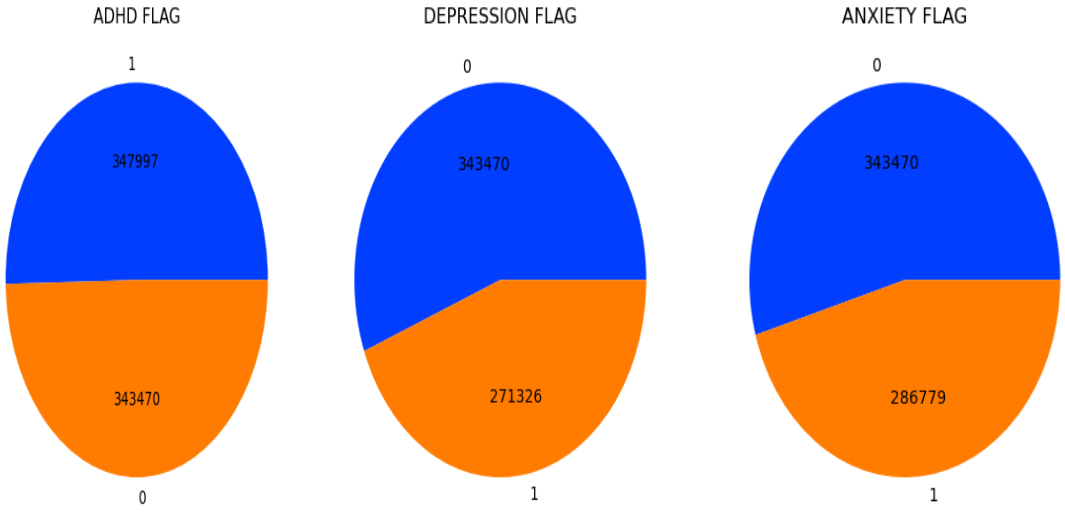
The supplemental data was aggregated at the district level to get a score on state level. This was finally joined to the original dataset from SAMHSA for forming a flat table. Rows were omitted randomly for maintaining a balance in the number of records for each of our four categories: ADHD Disorder, ANXIETY Disorder, DEPRESSION disorder, and no disorder. Finally, a total of ~1.1 million rows and 96 variables were retained.

Significant variables:

The initial thought was to use a multiclass classification approach to predict a client to have either ADHD, ANXIETY, DEPRESSION, Or a combination of two or more, or none of the above disorders. However, due to performance constraints and considering the business objective, 3 separate models to predict the corresponding disorder each for ADHD, ANXIETY, and DEPRESSION, were applied. Since a minor can

have one or more mental health disorders, it is important to focus on getting an accurate model for predicting each of the three mental health disorders in question.

To identify the most important variables that have a statistically significant impact on ADHD, DEPRESSION, and ANXIETY, a combination of T-test and Chi-square test at 1% significance threshold was applied to continuous and categorical variables respectively to ensure the best possible accuracy. Finally, a set of models were trained for predicting ADHD, ANXIETY, and DEPRESSION. A set of ~690K records and 86 variables were used to predict ADHD. A set of ~640K records and 86 variables were used to predict DEPRESSION. A set of ~630K records and 86 variables were used to predict ANXIETY.



Data Limitations:

The Mental Health Client Level Data (MH-CLD) offers insightful information about mental health demand, but exhibits several limitations:

Data Scope: MH-CLD offers valuable insights into mental health demand but is limited in its scope. The data does not capture the total national mental health demand, and supplementary data is essential for a more comprehensive understanding. To address this limitation, efforts should be made to incorporate additional data sources to ensure a more representative and complete illustration of the mental health landscape.

Missing Data and Bias: The uneven distribution of missing mental health diagnoses within MH-CLD poses a risk of biased prevalence rates. To mitigate this issue, it is crucial to investigate the root causes of missing data and implement strategies to address them. This may involve enhancing outreach to underrepresented groups and ensuring that data is collected uniformly across different demographic categories.

Diagnostic Limitations: MH-CLD's allowance of up to three mental health diagnoses per individual introduces complexities and may not represent a complete enumeration of all diagnoses. This raises concerns about potential underrepresentation of mental health conditions. To enhance accuracy, efforts

should be made to revise the diagnostic criteria and capture a more comprehensive range of mental health conditions for individuals served. This may involve consulting mental health professionals to refine the diagnostic framework.

Facility Variations: MH-CLD's data compilation, influenced by state variations in licensing and funding, poses challenges in achieving a standardized and uniform dataset. To address this limitation, collaboration with relevant stakeholders, including state health departments, mental health facilities, and regulatory bodies, is recommended.

Modeling:

The dataset for predicting each Mental Health disorder was divided using an 80-20% stratified split for training and testing, respectively. Stratified splitting ensured that the distribution of records remained consistent in both the train and test samples, balancing the target variable in each. Additionally, to boost accuracy and robustness, a 4-way cross-validation split was applied during the training of each model. Various models suitable for classification tasks were evaluated to identify those with the best precision and recall values. For instance,

1. Logistic Regression
2. Decision Trees
3. Support Vector Machines
4. Random Forest Classifiers
5. Gradient Boosting Machines
6. Extreme Gradient Boosting Machines
7. Cat Boost

Model Training

To enhance the efficiency and to optimize the hyperparameters a Grid Search Cross Validation method for automatic hyperparameter tuning was applied. Cross-validation (CV) and GridSearchCV are super useful in training machine learning models, like Random Forest. Think of CV as a method to check how well the model can perform on different chunks of data. It is like giving the model mini tests at each step using different portions of data to see how well it can predict on the run while training the model for optimum accuracy. GridSearchCV is like a smart helper. It uses a set of predefined parameters to automatically train the given model with combinations of hyperparameters from the set (called hyperparameters) and sees which one gives the best results. For example, in a Random Forest, the model was trained using different numbers of trees or different criteria for splitting the trees and then the best combination was automatically picked. Using these tools helps the model learn patterns better and provide robust results.

Model Results

The two main performance metrics for this model are Precision and Recall. Since the cost of screening is irrelevant as all the data used to predict any disorder is readily available without the need for extra screening, the focal metric should be recall for disorders (Labelled as 1). A range of values from 53% Precision and 55% Recall in Logistic regression to 79% Precision and 80% Recall for random forest classifiers were obtained.

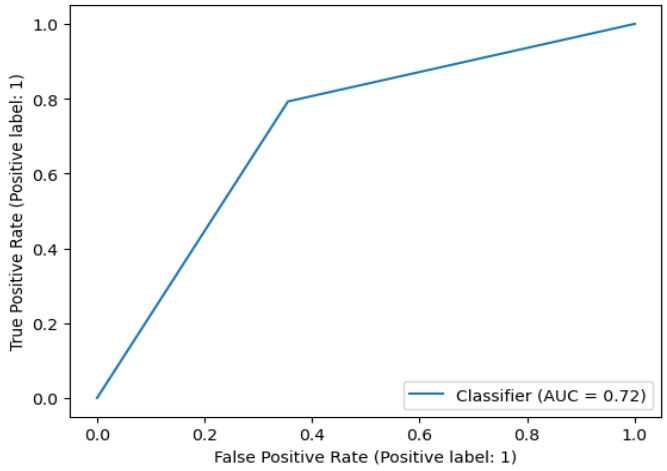
AUC (Area Under the cure):

AUC measures the ability of a model to distinguish between two classes, typically positive and negative. It reflects how well the model assigns higher scores to true positives and lower scores to true negatives.

Interpretation of AUC:

- 0.5 <= AUC < 0.7: Poor discrimination
- 0.7 <= AUC < 0.8: Acceptable discrimination
- 0.8 <= AUC < 0.9: Excellent discrimination
- AUC >= 0.9: Outstanding discrimination

AUC Curve for Random Forest Classifier predicting ADHD.



	Precision	Recall
0	0.76	0.64
1	0.69	0.79

The AUC value of 0.72 indicates that the model has acceptable discrimination ability, meaning it can distinguish between both classes better than random guessing.

Best parameters found: {'max_depth': 90, 'min_samples_split': 300, 'n_estimators': 1000}

The best hyperparameters for a Random Forest are:

'max_depth' of 90, 'min_samples_split' of 300, and 'n_estimators' of 1000 trees.

For Class 0

Precision of 76% suggests that the model is relatively accurate in identifying true negatives.

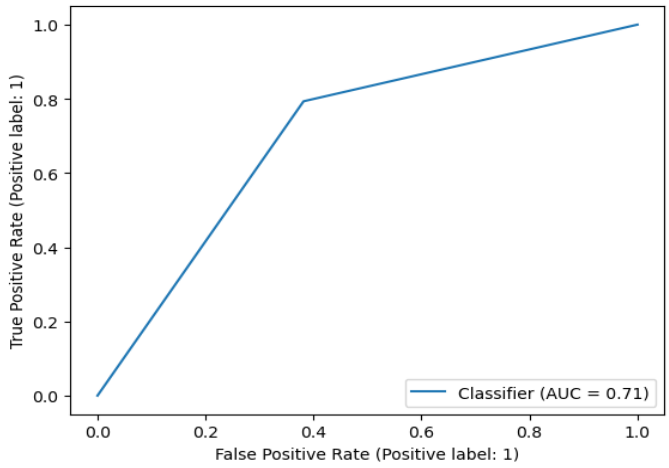
Recall of 64% indicates that the model correctly identifies only 64% of all actual instances.

For Class 1

Precision of 69% suggests that the model is good at identifying true positives.

Recall of 79% indicates that the model correctly identifies 79% of all actual instances.

AUC Curve for Logistic Regression predicting ADHD.



	Precision	Recall
0	0.75	0.62
1	0.68	0.79

The AUC value of 0.71 indicates that the model has acceptable discrimination ability, meaning it can distinguish between both classes better than random guessing.

Best parameters found: {'C': 1, 'penalty': 'l2'}

The best hyperparameters, 'C' set to 1, and 'penalty' as 'l2' (L2 regularization)

For Class 0

Precision of 75% suggests that the model is relatively accurate in identifying true negatives.

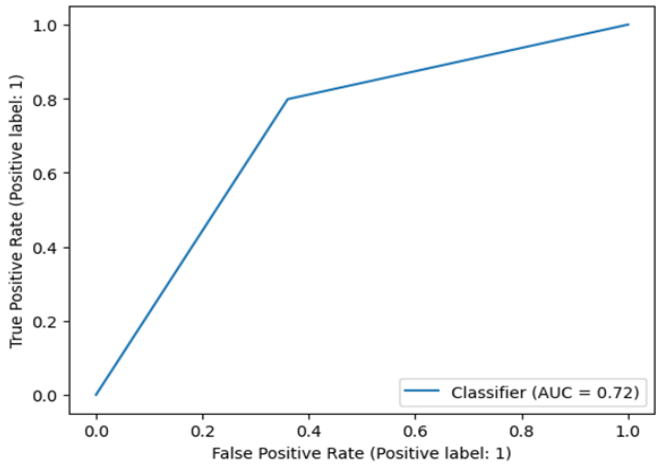
Recall of 62% indicates that the model correctly identifies only 62% of all actual instances.

For Class 1

Precision of 68% suggests that the model is good at identifying true positives.

Recall of 79% indicates that the model correctly identifies 79% of all actual instances.

AUC Curve for Gradient Boosting Machines predicting ADHD.



	Precision	Recall
0	0.76	0.64
1	0.69	0.80

The AUC value of 0.72 indicates that the model has acceptable discrimination ability, meaning it can distinguish between both classes better than random guessing.

Best parameters found: {'learning_rate': 0.1, 'max_depth': 7, 'n_estimators': 200}

The optimal hyperparameters for Gradient Boosting are 'learning_rate' of 0.1, 'max_depth' of 7, and 'n_estimators' of 200.

For Class 0

Precision of 76% suggests that the model is relatively accurate in identifying true negatives.

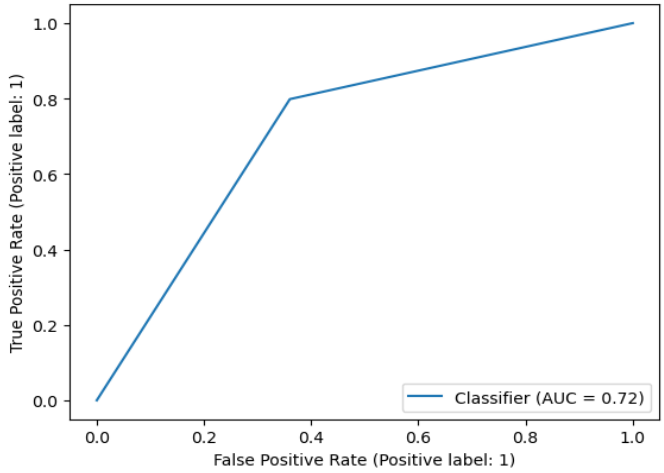
Recall of 64% indicates that the model correctly identifies only 64% of all actual instances.

For Class 1

Precision of 69% suggests that the model is good at identifying true positives.

Recall of 79% indicates that the model correctly identifies 79% of all actual instances.

AUC Curve for XGBoost predicting ADHD.



	Precision	Recall
0	0.76	0.64
1	0.69	0.80

The AUC value of 0.72 indicates that the model has acceptable discrimination ability, meaning it can distinguish between both classes better than random guessing.

For Class 0

Precision of 76% suggests that the model is relatively accurate in identifying true negatives.

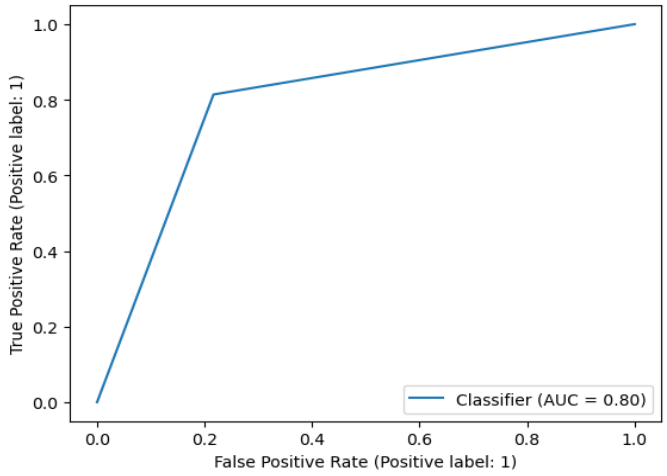
Recall of 64% indicates that the model correctly identifies only 64% of all actual instances.

For Class 1

Precision of 69% suggests that the model is good at identifying true positives.

Recall of 79% indicates that the model correctly identifies 79% of all actual instances.

AUC Curve for Random Forest classifier predicting DEPRESSION.



	Precision	Recall
0	0.84	0.78
1	0.75	0.81

An AUC of 0.8 means that your model has good discrimination ability. This means it can distinguish between two classes better than random guessing.

Best parameters found: {'max_depth': 50, 'min_samples_split': 300, 'n_estimators': 500}

The best hyperparameters for a Random Forest are:

'max_depth' of 90, 'min_samples_split' of 300, and 'n_estimators' of 1000 trees.

For Class 0

Precision of 84% suggests that the model is relatively accurate in identifying true negatives.

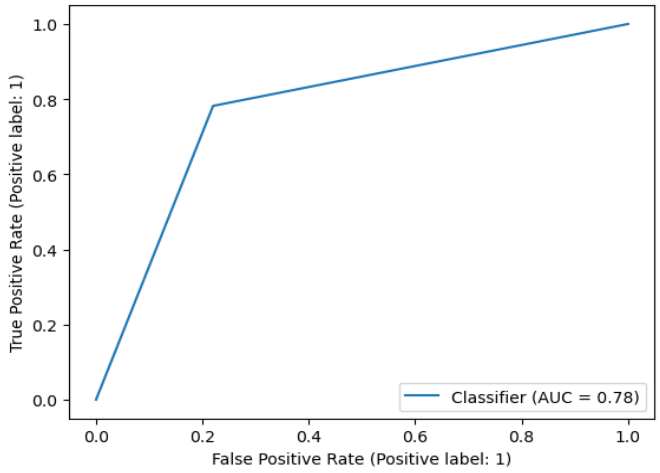
Recall of 78% indicates that the model correctly identifies only 78% of all actual instances.

For Class 1

Precision of 75% suggests that the model is good at identifying true positives.

Recall of 81% indicates that the model correctly identifies 81% of all actual instances.

AUC Curve for Logistic regression predicting Depression.



	Precision	Recall
0	0.82	0.78
1	0.74	0.78

The AUC value of 0.78 indicates that the model has acceptable discrimination ability, meaning it can distinguish between both classes better than random guessing.

Best parameters found: {'C': 10, 'penalty': 'l2'}

The best hyperparameters, 'C' set to 10, and 'penalty' as 'l2' (L2 regularization)

For Class 0

Precision of 82% suggests that the model is relatively accurate in identifying true negatives.

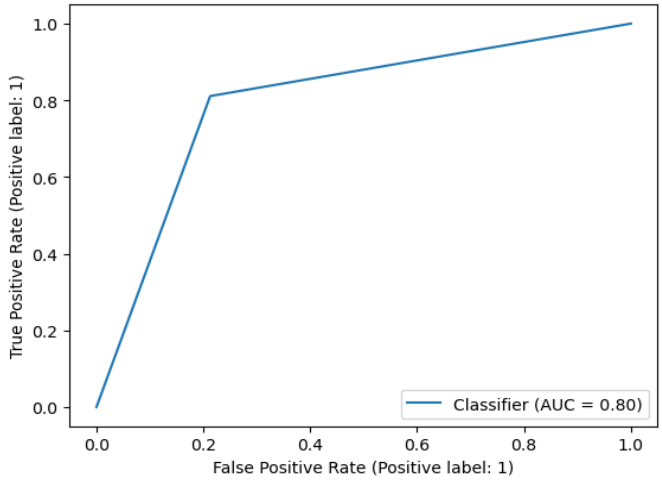
Recall of 78% indicates that the model correctly identifies only 78% of all actual instances.

For Class 1

Precision of 74% suggests that the model is good at identifying true positives.

Recall of 79% indicates that the model correctly identifies 78% of all actual instances.

AUC Curve for gradient boosting machines predicting Depression.



	Precision	Recall
0	0.84	0.79
1	0.75	0.81

An AUC of 0.8 means that your model has good discrimination ability. This means it can distinguish between two classes better than random guessing.

Best parameters found: {'learning_rate': 0.1, 'max_depth': 7, 'n_estimators': 200}

The optimal hyperparameters for Gradient Boosting are 'learning_rate' of 0.1, 'max_depth' of 7, and 'n_estimators' of 200.

For Class 0

Precision of 84% suggests that the model is relatively accurate in identifying true negatives.

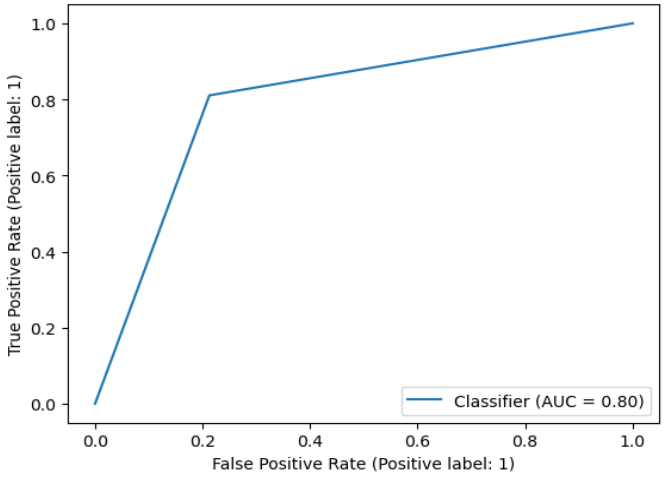
Recall of 79% indicates that the model correctly identifies only 64% of all actual instances.

For Class 1

Precision of 75% suggests that the model is good at identifying true positives.

Recall of 81% indicates that the model correctly identifies 81% of all actual instances.

AUC Curve for Xtreme Gradient Boosting machines predicting Depression.



	Precision	Recall
0	0.84	0.79
1	0.75	0.81

An AUC of 0.8 means that your model has good discrimination ability. This means it can distinguish between two classes better than random guessing.

For Class 0

Precision of 84% suggests that the model is relatively accurate in identifying true negatives.

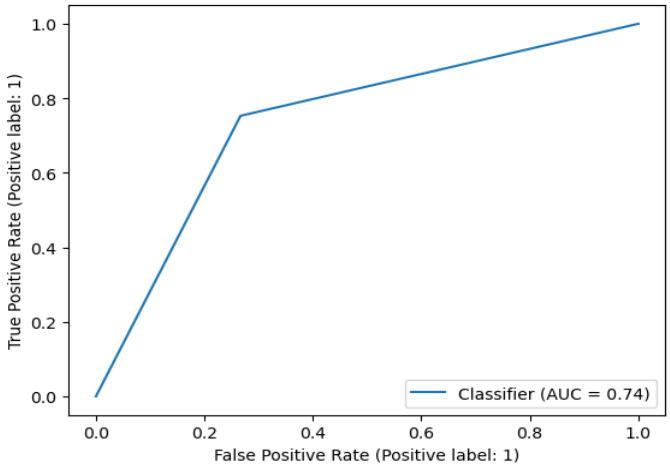
Recall of 79% indicates that the model correctly identifies only 79% of all actual instances.

For Class 1

Precision of 75% suggests that the model is good at identifying true positives.

Recall of 81% indicates that the model correctly identifies 81% of all actual instances.

AUC curve for Random Forest classifier predicting Anxiety.



	Precision	Recall
0	0.78	0.73
1	0.70	0.75

The AUC value of 0.74 indicates that the model has acceptable discrimination ability, meaning it can distinguish between both classes better than random guessing.

Best parameters found: {'max_depth': 30, 'min_samples_split': 300, 'n_estimators': 500}

The best hyperparameters for a Random Forest are:

'max_depth' of 30, 'min_samples_split' of 300, and 'n_estimators' of 500 trees.

For Class 0

Precision of 78% suggests that the model is relatively accurate in identifying true negatives.

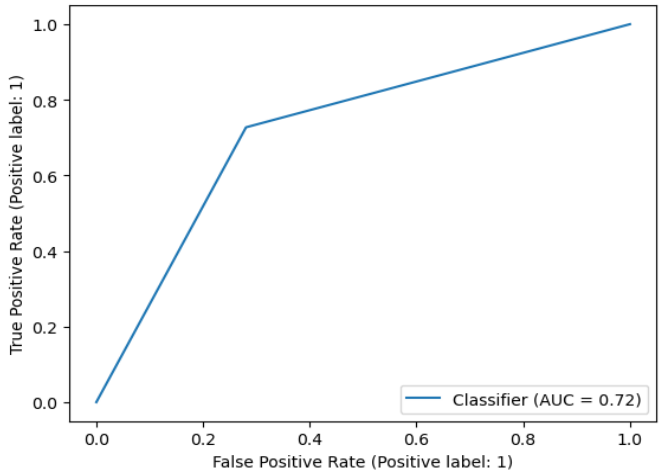
Recall of 73% indicates that the model correctly identifies only 73% of all actual instances.

For Class 1

Precision of 70% suggests that the model is good at identifying true positives.

Recall of 75% indicates that the model correctly identifies 75% of all actual instances.

AUC curve for logistic regression predicting Anxiety.



	Precision	Recall
0	0.76	0.72
1	0.68	0.73

The AUC value of 0.72 indicates that the model has acceptable discrimination ability, meaning it can distinguish between both classes better than random guessing.

Best parameters found: {'C': 10, 'penalty': 'l2'}

The best hyperparameters, 'C' set to 10, and 'penalty' as 'l2' (L2 regularization)

For Class 0

Precision of 76% suggests that the model is relatively accurate in identifying true negatives.

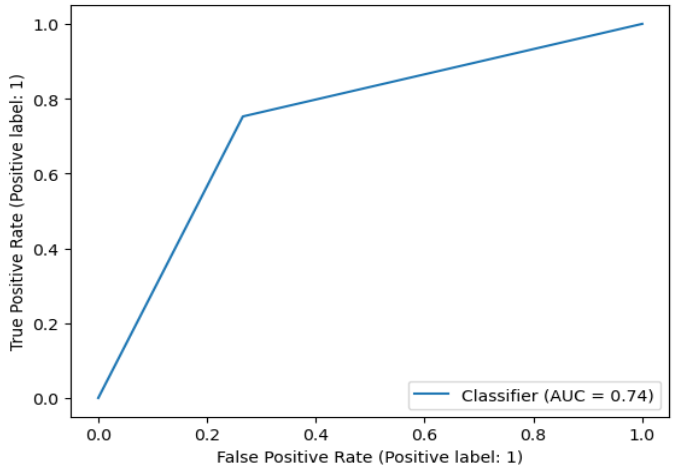
Recall of 72% indicates that the model correctly identifies only 64% of all actual instances.

For Class 1

Precision of 68% suggests that the model is good at identifying true positives.

Recall of 73% indicates that the model correctly identifies 73% of all actual instances.

AUC curve for Gradient Boosting Machines predicting Anxiety.



	Precision	Recall
0	0.78	0.73
1	0.70	0.75

The AUC value of 0.74 indicates that the model has acceptable discrimination ability, meaning it can distinguish between both classes better than random guessing.

Best parameters found: {'learning_rate': 0.1, 'max_depth': 5, 'n_estimators': 200}

The optimal hyperparameters for Gradient Boosting are 'learning_rate' of 0.1, 'max_depth' of 5, and 'n_estimators' of 200.

For Class 0

Precision of 78% suggests that the model is relatively accurate in identifying true negatives.

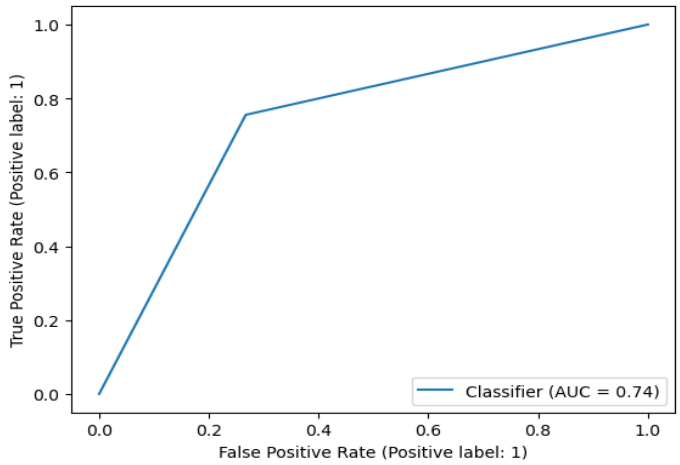
Recall of 73% indicates that the model correctly identifies only 73% of all actual instances.

For Class 1

Precision of 70% suggests that the model is good at identifying true positives.

Recall of 75% indicates that the model correctly identifies 75% of all actual instances.

AUC curve for Xtreme gradient boosting machines predicting Anxiety.



	Precision	Recall
0	0.78	0.73
1	0.70	0.76

The AUC value of 0.74 indicates that the model has acceptable discrimination ability, meaning it can distinguish between both classes better than random guessing.

For Class 0

Precision of 78% suggests that the model is relatively accurate in identifying true negatives.

Recall of 73% indicates that the model correctly identifies only 73% of all actual instances.

For Class 1

Precision of 70% suggests that the model is good at identifying true positives.

Recall of 76% indicates that the model correctly identifies 76% of all actual instances.

Future Scope for improvements: Reinforcement Learning

The idea of continuous training for machine learning models, in response to new data, offers significant advantages over traditional static models. This approach allows the model to constantly adapt and improve its performance as it encounters new information.

Benefits:

1. Allowing the model to learn from new experiences while leveraging its existing data. This can enhance the efficiency of learning and prevent the model from becoming too focused on past experiences.

2.Continuous training can enhance the model's ability to generalize to unseen data, leading to better performance on real-world applications.

References

Centers for Disease Control and Prevention. (2023, March 8). *Data and statistics on children's Mental Health*. Centers for Disease Control and Prevention. https://www.cdc.gov/childrensmentalhealth/data.html#ref

Frequently asked questions. SAMHSA. (n.d.). https://www.samhsa.gov/about-us/frequently-asked-questions#:~:text=Established%20by%20Congress%20in%201992,behavioral%20health%20of%20the%20nation.

Mental health client-level data (MH-CLD) client-level mental health data. Mental Health Client-Level Data 2020 (MH-CLD-2020-DS0001) | SAMHDA. (n.d.). https://www.datafiles.samhsa.gov/dataset/mental-health-client-level-data-2020-mh-cld-2020-ds0001

Prb. (n.d.). *Anxiety and depression increase among U.S. Youth, 2022 Kids Counts Data Book shows*. PRB. https://www.prb.org/resources/anxiety-and-depression-increase-among-u-s-youth-2022-kids-counts-data-book-shows/

Schlack, R., Peerenboom, N., Neuperdt, L., Junker, S., & Beyer, A.-K. (2021, December 8). *The effects of mental health problems in childhood and adolescence in young adults: Results of the Kiggs cohort*. Journal of health monitoring. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8734087/#:~:text=Mental%20health%20problems%20during%20childhood,quality%20of%20life%20as%20adults

The Children's Hospital of Philadelphia. (2019, September 25). *Youth suicide prevention, intervention, and research center*. Children's Hospital of Philadelphia. https://www.chop.edu/centers-programs/youth-suicide-prevention-intervention-and-research-center